EDUCATIONAL MEASUREMENTS
AND THE
CLASSROOM TEACHER

CENTURY EDUCATION SERIES

EDITED BY

CHARLES E. CHADSEY

PRINCIPLES OF TEACHING HIGH SCHOOL PUPILS. By Hubert Wilbur Nutt, Ohio Wesleyan University.

PSYCHOLOGY AND THE SCHOOL. By Edward Herbert Cameron, Ph.D., University of Illinois.

THE TEACHER'S TECHNIQUE. By Charles Elmer Holley, Ph.D., James Millikin University.

THE AMERICAN ELEMENTARY SCHOOL. By John Louis Horn, Ed.D., Mills College.

MODERN METHODS AND THE ELEMENTARY CURRICULUM. By Claude A. Phillips, Ph.D., University of Missouri.

THE EDUCATION OF EXCEPTIONAL CHILDREN. By John Louis Horn, Ed.D., Mills College.

EDUCATIONAL MEASUREMENTS AND THE CLASSROOM TEACHER. By R. H. Jordan, Ph.D., Cornell University, and A. R. Gilliland, Ph.D., Northwestern University.

PERSONNEL PROBLEMS OF THE TEACHER AND SUPERINTENDENT. By E. E. Lewis, Superintendent of Schools, Flint, Michigan.

THE TEACHING OF HIGH SCHOOL SUBJECTS. By William A. Millis, President of Hanover College, and Harriet H. Millis.

THE CONTENT AND METHODS OF INDUSTRIAL ARTS. By Samuel J. Vaughn, Hardin Junior College, and Arthur B. Mays, University of Illinois.

A MANUAL FOR SCHOOL OFFICERS. By W. N. Anderson, Los Angeles, Cal.

Other volumes to be arranged.

Educational Measurements
AND THE
Classroom Teacher

BY

A. R. GILLILAND, PH.D.

PROFESSOR OF PSYCHOLOGY, NORTHWESTERN UNIVERSITY

AND

R. H. JORDAN, PH.D.

PROFESSOR OF EDUCATION, CORNELL UNIVERSITY

THE CENTURY CO.

New York & London

1924

ACKNOWLEDGMENTS

It is impossible to give due credit in a text such as this to all the sources from which material has been obtained. Some material has been copied directly from various sources. For this material we have endeavored to give due credit. Other material has been included in a modified form, while some has grown out of mere suggestions from different sources. For material from these latter sources credit has not always been acknowledged. In a subject such as educational measurements the fund of material is already so great that original sources are often very difficult to discover. Despite these facts the authors are very desirous and have attempted to give credit where credit is due.

In addition to acknowledgments to publishers of textbooks from which material has been quoted special acknowledgments should be made to the many publishers and authors of the tests and scales from which quotations and illustrations have been freely taken. Chief among these are: The Public School Publishing Co., World Book Co., Bureau of Publications, Teachers College, Columbia University, Russell Sage Foundation, Mr. S. A. Courtis, Dr. M. J. Van Wagenen, and Dr. L. W. Pressey.

The authors are especially indebted to Miss Margaret Gessford of Washington, D. C., who painstakingly read the original manuscript and gave many valuable suggestions for its revision.

CONTENTS

INTRODUCTION

In presenting an addition to the list of books treating of the general subject of educational measurements, some definite justification must be offered. The authors of the present text feel that there is a twofold need which has been but partially met by the otherwise admirable books now before the public. The first is the need for a text which may be used as a hand-book for guidance of the teacher in service; the second is the need for a classroom text adapted to the use of prospective classroom teachers.

As to the first, there seems to be a haziness in the minds of many classroom teachers in both elementary and secondary schools as to the use they may make of tests and scales in their own work. They are too often of the opinion that tests are primarily supervisory instruments and so too difficult in operation and too abstruse in interpretation to be of any real aid to the individual teacher. A certain emphasis then should be put upon the fact that achievement tests are valuable instruments for the teacher to understand and use, independently of, or in co-operation with, the supervisor. The teacher should understand that she may handle her own work more intelligently, with more successful results, in proportion as she makes the greater use of proper measures of her effort.

Again, there has been a confusion in the terminology

of the tests, which has given rise to many false ideas among the rank and file of the teaching profession. The expression "psychological tests" has been used as a blanket phrase to cover every sort of measurement, and many teachers have not learned to discriminate between various types of psychological investigation. Comparatively few teachers, to judge from reactions in representative summer school groups, can explain accurately the difference between intelligence tests and achievement scales. Thus an unfortunate situation holds which must be cleared up in order that the work of the clinician and supervisor may be done with assurance of understanding and resultant sympathy from the body of classroom instructors.

The second great need refers to two aspects of the work in classes in education in normal school and college which deserve a rather different emphasis than has hitherto been given. On the one hand is the class of students preparing primarily for classroom teaching, who wish to study the technique and meaning of achievement tests at the same time that they are given practice in their use and evaluation. In the single semester frequently given to this work, they do not wish to include a consideration of the intelligence test or other psychological study, and so need a manual with the major emphasis upon the use of the achievement measure. The study of the intelligence test will come in a separate course, as will the factors of supervision and administration as related to measurement.

On the other hand is a group frequently met with in college departments of education, made up of undergraduates who are concerned more with general method and

subject matter of instruction than in the psychological phases of the teaching process. Unfortunately there are yet found a great many of this sort of students; and the teacher of general methods of instruction finds that they have no idea of the meaning or use of the tests, and that frequently they do not have a place in their undergraduate programs for such special courses. There is a place for a manual which may accompany such courses in general method, to give this knowledge of the various tests and their uses, and so to prevent the student from going into the work entirely ignorant of the great possibilities of such instruments. The authors have had the thought of working out a text-book which may be adapted to both types of undergraduate classes.

The present volume, then, is an attempt to meet this twofold need; in order to meet the varying objectives, the idea has been kept steadily in mind of presenting the important concepts and methods of application of tests and scales in as simple and non-technical language as is consistent with sound practice in their employment. The order of topics and general arrangement have been planned from this point of view. It is hoped that the materials are so presented that the book will find its place as a manual for the teacher in service and a suitable part of the reading circle on the one hand, and as a classroom text for the normal school, college, and school of education on the other.

EDUCATIONAL MEASUREMENTS
AND THE
CLASSROOM TEACHER

EDUCATIONAL MEASUREMENTS AND THE CLASSROOM TEACHER

CHAPTER I

REASONS FOR EDUCATIONAL MEASUREMENTS

A district superintendent of schools in one of our most progressive states asked this question a short time ago: "Tell me, is n't this matter of the educational test and scale dying out? I don't hear so much said about it recently, and I am wondering if it has n't proved to be another passing fad?" In the light of similar remarks from other sources, the thought persists that after all, the real need for such measurements as have been developed in the past decade is not generally understood by teachers and superintendents, especially in the village and rural schools, and that the best way to open our discussion is to review briefly the reasons for attempting to make better educational measures than we have had in the past.

Three questions arise which must be answered in order to make these reasons clear: First, what need has the teacher for any concrete standards for measuring the work of the classroom? Second, why are the methods of the past not adequate? Third, is there any real progress being made toward a solution of this problem, which may

afford relief to the class teacher? These three questions are closely related, and will be taken up in order.

FIRST: The teacher needs some sort of concrete standard of measurement in at least three inevitable relations: (a) the need in relation to the public, particularly as represented by the parents of the pupils; (b) the need in relation to the superintendent of schools, the supervisor, the principal, or other officer in immediate charge of the teachers; (c) the need for standards in the relation with the pupils themselves, involving the teacher's own guidance in the classroom.

Need of standards in relation to the public—The teacher must have some sort of means for acquainting the parent with the attainment of his child. Fathers and mothers expect this report as a matter of course, and if it is not forthcoming, insist upon it. Thus one reason for the failure of attempts to do away with school marks and marking systems has been the refusal of parents to agree that measures of progress are unnecessary. Doubtless the typical parent has been improperly educated in the application of such standards, especially to his own children, but this makes all the more important a kind of mark which will carry with it a clear idea of the reasons for the child's success or failure. Very few parents are content with a system which designates simply "Passed" or "Failed" as the teacher's verdict on the work of the month or semester. Parents generally wish to know something of the relative positions of their children; whether the children are working somewhere near their mental capacity; whether the child is diligent, whether he tries to succeed, and something of the general character of the attempt.

The parent is not the only person interested in the work of the teacher. The mother's clubs, and other women's organizations, the board of education, the rotary clubs, are all bodies which are interested in the work of the individual teacher as well as of the school system in general. The teacher is often asked to appear before some one or other of these bodies, sometimes to explain the work of the class for the information of the public, sometimes on the defensive, to meet some criticism, sometimes as a member of the body, to extend the influence of the school. In such relations, hastily explained and readily understood standards are most essential to give point and emphasis to her report. If she can show by definite comparisons that her work in such a subject as arithmetic or history is better than that of the United States standard for the subject, she scores a distinct triumph and wins approval for her school and her city; particularly is such a result a happy one, if her teaching has been criticized or attacked; for when she has a standard with which to compare her work specifically, and to demonstrate without cavil her success, she can silence critics in the most satisfactory way.

Need of standards in relations with supervisory officers—The work of the supervisor is to improve the results of the classroom teacher. This holds for the relation of the superintendent, the principal, the special supervisor, to the teacher. Wise supervision is directed to the strengthening of the teacher, and to capitalizing her strong points, not to the criticism of the instructor's weaknesses, or to an attempt to bring out her failures in bold relief. Her successes are to become habitual, her failures to become negligible. But in the past, there has

been much difficulty in the way. Teachers have claimed successes which the supervisor could not recognize, for lack of some easily applicable measure of success. Supervisors have set up standards for judgment which the teacher could not understand, and, indeed, the supervisor has frequently not deigned to explain to the teacher the standard by which the judgment was made. One reason for this has been that the standard existed only in the mind of the superintendent, and had not been defined in concrete terms in his own consciousness. He even might go so far as to say that his standard was indefinable—that good teaching and bad teaching were to be "sensed," but that the difference could not be expressed in definite terms. Under such supervision, the teacher was largely helpless; she knew whether her work was approved or not, but she was not conscious of any standard by which this result was reached. So the need has been felt very definitely for standards by which the ideas of the supervisor might be passed on to the teacher in terms of objectives to be striven for by the teacher with a clear knowledge of the goal to be attempted; and by which there could be a clear understanding by both parties of the final decision as to success or failure to reach this goal. Such standards are necessary if there is to be the sympathy between supervisors and teachers necessary to the best results. This sympathy has not obtained in the past in a large number of our communities; and there has grown up an antagonism between the supervisors and the teachers which has retarded the work of the schools. If concrete measurements are used, much of this antagonism would disappear. The supervisor

would not only be welcome in the classroom, but would be sought after, and invited.

Need for standards in relation to pupils—Even though the teacher did not have to justify her work to parents or to supervisors, she would nevertheless be faced with the necessity of satisfying herself as to the efficacy of her methods, and as to her success in handling individual pupils. When each pupil has a specific goal set for him to approach, he will work much more intelligently and frequently more willingly than otherwise. When he understands the reason for assignment of a recitation mark, or a semester evaluation, he will be less likely to charge his success or failure to the whim or partiality of the instructor. If this standard is sufficiently clear, he will be able to assign his own mark without the evaluation of the teacher being involved at all. He will also understand his mark in relation to the marks received by other pupils, so that he will answer for himself the question "Why did Beatrice get 'G' on her report when I do just as good work as she, and I got only 'F'?" Thus the better the standard, the more likely the teacher is to have coöperation and zealous effort from the pupils.

Again, the teacher should have standards which will adequately indicate the real differences between individual pupils; for the present emphasis upon study of individual differences between children demands a measure of these varying capacities which is definitely objective.

The matter of a final determination of the pupil ready for promotion, or of the child to be kept back, demands a standard which will be based upon performance, not

opinion, and which will further be of real assistance to the teacher in deciding this most important question.

More than that, the instructor frequently wishes to know the answer to the question: "Have I gotten the results that I ought to have obtained? Is my class up to standard?" Many a teacher has worked a lifetime actually without knowing whether her work has been really good, or really bad, superior, or mediocre. Surely every teacher wants some definite information as to the true quality of her work. And she would much prefer to find this out for herself, than to have it discovered for her by a supervisor or inspector. If she can have proper standards in her hands, she can do this. The need is very apparent.

A comparison of the results of tests in her various classes with the norms made by classes generally over the country, as shown by the measurement of large numbers of children of all types, will give her a basis with which she can compare the result of the same, or equally difficult tests given at an interval of a month, or longer, and thus have a definite measure of the effect of her teaching. This will enable her to measure progress of the pupils, and to know whether she has done as good a piece of teaching as she had supposed. She may find that she has really been succeeding, where she could not realize that progress was being made; or she may find that the marked improvement she had noticed was not so great as she should have obtained. But in time she will be able to estimate her own proficiency more capably than was ever before possible.

SECOND: The methods of the past are not adequate to meet these various needs because they have not been

based upon sufficiently sound scientific or even pedagogic methods. Let us examine the methods most frequently used. They have been (a) examinations; (b) marks based upon recitations and tests; (c) opinions of teachers and supervisors.

Inadequacy of examinations—The examinations as ordinarily given in the past have been faulty in one or all of the following particulars: (1) they have been constructed without a clear understanding of their purpose. Sometimes the purpose is to test the pupil, sometimes to test the teacher through the pupil. Neither of these has been made entirely clear in many tests. If to test the pupil, is the idea to find out how well he has done certain assigned work, or to find out how much he knows about a subject which he has been studying? He may fail on the latter, but pass on the former. Is the examination to test memory, or reasoning power, or retention of information, or reproduction of ideas, or what? When examinations have been set by persons not in control of a specific class, they have frequently failed of their purpose because the teacher has not understood the purpose in the mind of the examiner, and the pupils have consequently not been trained to meet the particular idea suggested. This has been unfair to teacher and pupil alike.

(2) They have not been constructed so as to make possible an accurate rating when corrected. This has reference to the values attached to the various parts of the paper set. In fact, there has never been any real agreement among teachers as to proper standards for rating the various parts of a paper. If ten questions are set on a paper, what should be the allotted value of each? If this question is raised with reference to a

specific paper in mathematics, or history, or English, or any other subject, and the answer requested from a group of experienced teachers of the subject, there has never been found any agreement as to the evaluation of the various questions. When Question One is valued at ten points by one teacher, and at eight points by a second, and at three points by a third, all on a scale of 100, it will be seen that the pupil failing to answer this question only, will have a mark varying from 90 to 97 for his paper. Surely examinations are inadequate standards when there is no more accuracy in their framing than this!

(3) They are not corrected accurately even when made properly. Whenever papers have been sent to groups of teachers to be corrected according to given weights, but without further instructions, the variations have been so marked as to reveal the greatest inaccuracies in marking generally. This holds not only for such subjects as composition, which is supposed to be largely evaluated by judgment, but in mathematics, in which judgment in marking is not presumed to be an important factor. If Example One on an arithmetic or an algebra paper is weighted eight points for a perfect answer, shall one point, or all eight, be deducted for a single error in computation, the principle of the work being correct? Teachers are not by any means agreed on this point, and so a paper with this single error may be marked anywhere from 93 to 99 points on a scale of 100. One does not have to go any further to show the lack of agreement resulting from this failure to evaluate errors on a common basis.

A more astonishing fact has been brought out by experiment,[1] in that papers corrected by groups of teachers have been submitted to them after a considerable lapse of time, for re-correction, and it has been demonstrated that the same teachers will not correct the same papers twice alike. The variation of markings given under these circumstances is indeed surprising.[2]

Enough has been said to show that the examination in itself has been a most inadequate standard. Probably the most satisfactory sort of paper is that which is set by the teacher herself to test her own class; but even this is not always suitable; how many times we have heard teachers say: "I gave my class an examination today, and they did so badly. I thought they knew the work perfectly, but they made terrible mistakes," or, "I made my examination too easy, I think; everybody passed."

It is presumed that the experienced teacher has a definite plan in mind in examining, by which the paper is set with direct relation to the teaching, and with definite purposes in mind. But remarks such as those quoted do not indicate any such idea. In fact, until recently the matter of making examination papers has never been made the objective in educational classes, and so very few teachers recognize that there is a really scientific basis for making such papers.

Inadequacy of marks based on recitations and tests—Many teachers have realized that examinations given at the close of work periods, as at the close of the semester, are not accurate measures of the pupil's ability,

[1] Daniel Starch, *Educational Measurements.*
[2] See Bibliography at end of chapter.

and so have either discarded the formal semester examination, or modified it by marks based upon the daily recitation and the quiz. Without doubt, such marks are more accurate measures than single examinations; but they are inadequate as really accurate standards for evaluation of the pupil's achievement. There are several reasons for this that immediately suggest themselves:

(1) They are fragmentary. Unless the teacher is so great a slave to the record book that every pupil response is recorded, and this of course means inferior teaching, the record is after all only a partial one, and so is open to dispute as to accuracy.

(2) They are not evenly balanced. The number of pupil responses is not the same for all; one great criticism of teachers by pupils is that some children are called upon in class much more frequently than others. The pupil responses are not of equal difficulty; another criticism often made is: "She gives me all the hard questions."

(3) Where the written quiz is used, the same objections arise as for the more formal examination.

(4) Just as there is great variation in marking papers, so there is great variation in evaluating oral responses.

(5) The art of questioning, and purposes of the question, are not much better understood by many teachers, than is the science of making examination papers. This again makes the recitation an unsafe guide for marking.

To be sure, the average of a number of marks may conceivably be more accurate than any single mark, either of recitation or examination. But the teacher can not safely rely upon an average to correct single inaccuracies, for the average of unreliable scores is more reliable than

the individual score only when the errors are equally distributed on both sides of the true measure. This is very unlikely to happen in assigning class marks. So the law of averages can not be invoked to correct marking errors.

Inadequacy of opinions of teachers and supervisors—Results of examinations and averages of classroom marks as bases for evaluating work of teacher, class, or individual pupil have for a long time been recognized by keen observers as inadequate measures. But the conclusion reached during a period of many years was that no more adequate measures could be made, and that the safest guide in applying these measures was the experience and good judgment of the teacher and superintendent. Accordingly in many cases where the faulty examination produced the failing pupil, the teacher substituted her judgment as to the pupil's ability, and passed him to the next school grade, regardless of his seeming failure. Even where the class record was not good, the superintendent declared that since the teacher had used methods which he thoroughly approved from his own experience, he would promote the teacher despite her apparent failure. Doubtless a teacher of long experience who had developed unusual skill in diagnosis might employ such methods with comparative success. The judgment of a good superintendent in evaluating work of teachers and classes is not to be thrown aside as of no value. But while a few persons unquestionably have such qualities, the rank and file of teachers and supervisors lack experience. Without experience on which to base judgment, keenness in diagnosis is not likely to command much respect, where it is based upon one's opinion

solely. Such judgments are spoken of generally as subjective, that is, they are dependent upon the idea originating within the mind of the person giving the opinion, and upon his own individual conclusion, reached by considering data developed from his own experience or his own personal tests. A subjective judgment can never have the force of an objective conclusion, for the objective idea is based not upon one's own experience, but upon data obtained by methods set as a result of combined experiences and judgments of others, and is more concrete in form and substance than the subjective thought. The tendency of the subjective conclusion is to disregard objective data; the objective conclusion is directly dependent upon objective data. The subjective judgment is therefore always open to criticism, to doubt, and to successful attack. The objective conclusion is in just the reverse position; based upon data scientifically collected and standardized, it is impregnable.

THIRD: As a result of the experiments and studies made during the past fifteen years, the third great question may be answered categorically.

Progress actually being made—Great progress has been made, and is being made, to solve the problem of proper measures of educational attainment. Instead of dying out, or losing force, as our district superintendent thought, the movement is becoming so generally accepted in principle, that it has passed the stage of discussion, and for this very reason is not occupying the space in controversial literature that it did ten years ago. At that time, the question was: "Is it possible to measure any educational achievement objectively?" Now the only question is: "Is there any educational achievement which

can not be measured objectively?" The wide-awake teacher may now find it possible to measure the greater part of elementary school attainment and a large part of the secondary school curriculum by well standardized objective tests or scales. Every year more and better measures are being produced and hundreds of men and women are working to perfect those already existing, and to extend the field of application still further.

Chapter II will discuss the nature of these measures in the light of the inadequacies of traditional methods, and will classify the various types of measures now in use.

REFERENCES TO AUTHORITIES ON INACCURACY OF TEACHERS' MARKS

Buckner, C. A., *Educational Diagnosis of Individual Pupils,* Teachers College, Columbia University, 1919.

Carter, R. E., "Correlation of Elementary Schools and High Schools," *Elementary School Teacher,* Vol. 12, p. 109.

Comin, R., "Teachers' Estimates of the Ability of Pupils," *School and Society,* Vol. 3, p. 67.

Inglis, Alexander, "Variability of Judgments in Equalizing Values in Grading," *Educational Administration and Supervision,* Vol. 2, p. 25.

Johnson, F. W., "A Study of High School Grades," *School Review,* Vol. 19, p. 13.

Kelly, F. J., "Teachers' Marks," *Teachers College Contributions to Education,* No. 66.

Monroe, DeVoss and Kelly, *Educational Tests and Measurements,* Houghton Mifflin Co., Chap. 1.

Starch, Daniel, *Educational Measurements,* Macmillan Co. Chap. 2.

Starch and Elliott, "Reliability of Grading High School Work in English," *School Review,* Vol. 20, p. 442; "Reliability of Grading Work in Mathematics," *School Review,* Vol. 21, p. 254; "Reliability of Grading Work in History," *School Review,* Vol. 21, p. 676.

CHAPTER II

WHAT CONSTITUTES A STANDARD MEASURE

The discussion of Chapter I has set before us very definitely the needs for measures of classroom work which will be free from any suspicion of unfairness in construction or application, which may be relied upon to tell us the exact truth about the progress of our pupils, and our own success or failure in presenting our work, and which are quite free from the whim or prejudice of either teacher or superintendent. To meet these needs a number of different sorts of measures have been devised. As yet, although marvelous progress has been made, the need has not been entirely met; but enough has been done to show that the making of standard measures is possible, and that although absolute perfection has not resulted, the measures now in use are so far in advance of the imperfect methods of tradition, that no teacher is justified in remaining in ignorance of their scope or of their applications.

This chapter will be devoted to a discussion of the factors involved in making such tests of achievement, and of the various classes of tests and measures available. This discussion will involve three phases: (a) Differentiations from traditional methods; (b) necessary specifications; (c) types of measures resulting.

Differentiations from traditional measures—To be en-

tirely satisfactory, our new measures must be free from the objections suggested in Chapter I as holding for the methods of the nineteenth and previous centuries. They must avoid the vagueness and indefiniteness of the old-fashioned test; they must be based upon a definite knowledge of the element to be tested, and must be framed to test that element, particularly. Each part of the test must have a definite value, ascertained by scientific methods. No part of the measure can be left to determination by opinion of the examiner, either in its origin or its application. So far as humanly possible the subjective element must be eliminated.

Now these various requirements are at present possible of being met in large part, whereas in the nineteenth century the necessary means were largely lacking. Only in the past quarter century have educators understood how to apply to educational products the mathematical methods used for many years by biologists, astronomers, and other scientists. Thus the last twenty-five years have created a scientific side of education which is an entirely new development, and makes possible many things before thought beyond our reach.

Educational experiments are not now considered valuable unless they are carried out under scientific principles, and controlled and interpreted by scientific method. Our measures are therefore to be based upon scientific experiment and mathematical interpretation. If the element of judgment is involved, it must also be subject to scientific scrutiny, and unless it can meet such searching inquiry, must be discarded as worthless. So even where a subjective factor may be involved, it becomes in the light of scientific method, objective in its application.

We then feel justified in speaking of all measures, tests, and scales, evolved under scientific principles, as objective measures both in evolution and in application. No longer can the opinion of one single individual, or limited group of individuals, have weight in framing measures of scientific value.

In appearance, some of the new tests resemble exactly the old subjective examination, and the uninitiated teacher can not discover the value of the new over the old. The value lies in the fact that the new in all of its parts has been subjected to scientific scrutiny which has established the fact that it is free from the faults of the old; it is known definitely what it is intended to test; every part has a scientifically determined value, and so is not the result of individual opinion or whim; in correction of answers to the questions, nothing is left to opinion; every answer has been weighed beforehand, and provision made for evaluation of every possible variation from the precise response desired. The personal equation no longer holds in framing or correcting this new paper. It is truly objective.

Necessary specifications of the objective measure—In order to avoid the weaknesses of traditional measures, and to meet modern scientific requirements, measures of classroom achievement must now meet certain definite specifications. The most important of these are as follows:

1. **Definite aim in mind**—The aim of instruction in the subject tested must be clearly envisaged. In teaching Addition the aim may be to render the subject of instruction competent to react automatically to a certain situation; in other words, to drill him until he is letter perfect

in various number combinations. The measure which tests these automatic reactions is then the essential measure, and must be made with this aim in view. But if the aim is to teach reasoning in arithmetic, through the medium of combinations involving Addition only, the new aim must be in mind in testing the pupil or class. The measure will then vary definitely from the first. A measure of Addition can not be made to conform to our new standards, unless it is known what aim is involved in the teaching. This principle holds for all subjects in the curriculum. Both the research scholar in collecting materials for the making of the measure, and the classroom teacher in using the measure, are at fault unless they have this first essential, the aim of the subject, definitely in mind.

2. **Material must be representative**—After the aim is clearly in mind, there must next be a selection of representative material which will develop and expand this aim. The materials entering into any subject must be comprehensive enough to make the realization of the aim possible, without involving extraneous elements which confuse and complicate the result. This is a principle of pedagogy adhered to in teaching the subject, and the same principle must be kept in mind in making the measure of the subject. Right here it must be emphasized again that objective measures are to be used to evaluate the results of teaching, and that good pedagogy is essential to the making of the scale. A test or scale which involves material rejected by competent teachers on the ground of poor adaptation to good instruction certainly is a poor scale. The test which is so limited in the scope of material covered that it does not give opportunity to show

the result of broad-gauge teaching, is not adapted to show the possibilities of excellent classroom instruction. The material must be in every way representative. It must be comprehensive, properly selected in scope and difficulty, adapted to the best teaching method, and suited to the best realization of the aim of the subject.

3. **Quantitative elements and methods of measuring them**—That subject matter is best suited to measurement by objective scales or tests which contains the greatest quantity of objective elements; the earliest objections to objective measures were based upon the statement that all teaching is so largely qualitative that no exact measures were possible. The objectors said: "How can one measure appreciation of a beautiful poem?" They felt that such matter of instruction held but few quantitative elements, and so was unsuited to measurement.

Obviously it is easier to make a measure for quantity rather than quality. So we feel surer of the results of a measure which deals with quantities, and measures of proficiency in spelling, of fundamental computations in arithmetic and of descriptive data in geography are accepted with but little hesitation. Measures of proficiency in penmanship, English composition, interpretation of history, are not so readily received as valid. But as the study of measurement has gone on, the amazing thing is that the quality of the most intangible matter is shown to have certain quantitative elements by which degrees of excellence in mastery are disclosed.

The appreciation of a poem can be measured by skilful questioning to a point where the teacher of literature feels quite justified in putting down a subjective mark

to indicate the relative interest, enthusiasm, esthetic understanding, and other factors of appreciation shown by different members of the class. So the measure must be based upon these factors of quantity, and the maker of the measure shows his ability by his success in separation of the quantitative from the other elements of the subject.

When there is a quantity of any thing, it is possible to measure this quantity. Our familiarity with ordinary measures is the very reason for our failing to understand the problem that confronted primitive man in devising proper measures. It is easy to imagine early men saying: "There is certainly such a thing as heat. But it is subjective, a thing of the senses, and varies in subjective sensation with each individual, according to his susceptibility. No measure can be found to scale this intangible sensation!" In the light of such reactions, how marvelous does the thermometer become! So it is with the products of the classroom. Our methods of measurement are based upon the same principles that were used by scientists in devising methods for measuring intensity of light, electric current, flow of gases, and such intangible things.

The arbitrary zero point on the thermometer is suggestive of a passing point or a failing point also chosen arbitrarily on our school product scale. But educational measurements in general are made on scientific principles which rule out such arbitrary assumptions; rather are the division points based on actual achievements of large numbers of individuals. So our methods are the outgrowth of experiment with, and trial of proposed

measures to determine those factors which can be scaled in relatively definite ways, with elimination of elements which do not lend themselves to such specific treatment.

A varied sort of measure results; we may measure reaction to certain stimuli on the part of the pupil, as when he is asked to perform certain tasks which are then corrected according to fixed standards; we may compare specimens of his work done under normal conditions with models containing the same or similar elements; we may evaluate oral responses; we may use his written reactions; we may measure physical as well as psychical factors, especially in connection with manual dexterity involved in such manipulatory subjects as penmanship, typing, telegraphy, and the like. In these last, a diagnosis of manipulation is necessary in order to make a satisfactory measure. In general, there is present an endeavor to diagnose the elements of mental activity involved in the various situations measured. It is this last factor which sometimes complicates the result of the work to an extent that makes care necessary in evaluating results, and frequently makes a re-examination of the individual important before passing judgment or drawing hasty conclusions. So the warning is given here, as it will be strongly emphasized hereafter, that the study of the individual child must always be made in the light of repeated and varied measures, rather than that his future be determined on the basis of a single measure or single set of measures.

With reference to the whole method of measuring quantitative elements, we must stress the fact that so far as possible measures must be based upon the possibility of isolating these elements from concomitant

factors, and that the validity of measures is affected principally by the extent to which this is accomplished.

4. **Standardization of results**—The chief difference between certain forms of examination, as those made up by examining boards, and objective measures of the sort described herein, is that the latter are standardized, while the former are purely experimental. The examination constructed by a large committee, each member of which actually participates in its composition, is not subjective, in the sense that it is the product of some one person's opinion, but as the result of collaboration of presumably competent judges, it is felt to be free from individual bias or prejudice. Yet at the best, it is untried in the sense that it represents only a consensus of opinion, and not a scientific result of experiment or computation.

Standardization of the test comes from the actual giving of the test to a large number of individuals of a given age or school grade, and a computation in the light of this experiment of the expectation to be reasonably held when it is to be given to any similar group. Thus a test which has been standardized has been proved of certain validity which makes it of tremendously more value than the test which is dependent for its value solely upon the opinion of the persons who devised it.

The standard of actual achievement made by groups of pupils in taking a test is called the norm of achievement. If a norm is desired to measure children of a given grade generally, this norm would of course be the result obtained by giving the test to all children in the United States in that particular grade, and then setting the norm as the amount of the test done successfully by the average, or presumably normal child, of the group. Of course,

it is impossible to give a test to all children; but fortunately the laws of mathematics demonstrate that if results are obtained from a large enough number of representative children in various parts of the country, they hold with equal validity for all children. This means that children must be tested in all sorts of social environments, with a proportionately larger number tested in schools of average social environment.

When this is done, the usual supposition is that a norm should be based upon the average performance of the group. There are many deviations in actual practice from this idea, dependent upon the type of norm desired, and the sort of test given. But unless other explanation is given, this is presumed to be the meaning of the norm. The number of cases needed to establish the norm can not be given with any degree of exactness, as again there enter the type of test, the sort of group designed to be tested, and many other elements. Usually it is sufficient to say the cases should be added until the inclusion of an additional hundred random cases causes practically no fluctuation in the general result.

The norm refers in psychological language to the result of actual achievement. An objective set before the teacher *for* achievement is known as a standard for achievement, and from this time on shall be so referred to in this book. The standard is therefore higher than the norm, and is to an extent theoretical in assumption and in application. The teacher must keep in mind that the norm is the measure which she is to use most commonly in connection with the standardized test.

Standardization of the test also is assisted by the placing of the results of the test in their proper order

according to the mathematical principles underlying the so-called normal probability curve.[1]

By way of summary, then, our measure must be made with distinct reference to the aim of the subject measured, the materials must be carefully selected with this aim in mind, quantitative elements must be selected susceptible to reasonably accurate measurement, and the resulting test or scale must be tried out until the results have a standard value by which all further results may be unequivocally placed.

Various kinds of measures—Most that has been written so far has referred more particularly to what is known as tests or scales to be used in connection with the classroom to determine the degree of attainment pupils have reached in their regular daily work in such subjects as arithmetic, language, spelling, Latin, history, etc. Such measures are known as achievement or attainment measures, to distinguish them from mental or intelligence tests, which do not have to do with school progress, but rather with native ability.

Measures of achievement are known as tests or scales; the idea underlying the term "test" is that a measure will be obtained which is made up of parts of equal difficulty, as the inches on a yardstick; the term "scale" is used for a measure in which the parts are not of equal difficulty, but are in an ascending series of difficulties, in a way comparable to a thermometer. Courtis' Arithmetic measures are tests; Woody's Arithmetic measures are scales. We speak of the Thorndike Alpha *Scale* for Reading, but of the Kansas Silent Reading *Tests*.

[1] The reader is referred to McCall's *How to Measure in Education* for the best statement of this method for the beginner in measurement.

Mental tests are quite different in idea from achievement tests. They are of two sorts, laboratory tests used by psychologists to test reactions of individuals to various stimuli, in order to determine their varying degrees of sensation, perception, and the like, and tests given to individuals or groups in the school room to determine their ability to answer questions or perform tasks of a general character not connected, at least directly, with school work of any sort.[2]

The test which has come into great prominence in recent years is the individual or group intelligence test which may be given independently of the laboratory, and has grown out of the work of the French psychologists Binet and Simon, and so known as the various revisions of the Binet-Simon tests, to make them applicable to American children. Of these, the best known is the Stanford Revision of the Binet-Simon tests, which is generally conceded to be the best of the individual tests. It is fully explained in Prof. Terman's *The Measurement of Intelligence.*

As a result of the mental testing done during the war, the group test has become very popular, and there are now a number of these, each having its coterie of advocates. These tests are to be given to pupils (or adults) in groups, as large as can be handled for lecture or recitation purposes. Some of the best known are the Army Alpha, the National, the Otis, the Haggerty, the Dearborn, the Myers, the Terman; but there are many more, perhaps equally as good as these mentioned.

[2] The tests of the psychological laboratory are well represented by the collection made by Prof. Whipple in his *Manual of Mental and Physical Tests.* The reader is referred to this excellent manual for a study of the various laboratory exercises and their significance.

There has been much controversy as to the definition of "intelligence," and as to what is really measured by these tests, so that one hesitates to enter upon controversial ground. There seems to be pretty good authority, however, for saying that in addition to whatever else the intelligence tests may measure, they determine with a good deal of accuracy the ability of pupils to do school tasks, in general. It may be said with confidence that a child with a high "intelligence quotient," "I. Q." (well above the average in his test score), has the ability to do better school work than the average pupil, and that the child with the low "I. Q." will lack ability to do as good work as the average, other things being equal. But the high I. Q. does not tell whether the pupil has the particular sort of ability which will make him more than average in arithmetic, or reading, or any specific subject. Nor does it measure those other qualities which make for success, as studiousness, industry, and the like. Therefore his high ability may not be reflected in actual accomplishment. On the other hand, the low I. Q. may be accompanied with such a high degree of application and perseverance that the child apparently handicapped may in the long run surpass the more favored person. A further statement will be made in this book regarding the whole matter of intelligence tests; but the main part of the discussion will deal with those measures which have to do with classroom achievement, and not with native ability.

The thought which should be left in one's mind from this presentation, is that the term "psychological test" has little meaning unless the sort of test used is indicated. In a good sized city recently the parents, and

some of the teachers, were much agitated because the superintendent of schools intended to promote the pupils at the close of the year on the basis of "pyschological tests." They interpreted this to mean that the intelligence test was to be used without reference to the pupil's actual advancement in his subject matter. Much criticism was indulged in openly, and some very loose talk resulted, in general harmful to the progress of the work, before it was discovered that the idea was to give achievement tests related intimately with the ground covered in class, instead of the usual final examination, and that the result of these tests was to be made the basis, but not the entire ground, for promotion. Intelligence tests were not to be given at all. Teachers must not jump to hasty conclusions with reference to any sort of measurement, until they know the exact facts, and they should take pains to get at the facts.

CHAPTER III

THE PRACTICAL USES OF EDUCATIONAL MEASURES IN THE CLASSROOM

So much has been written and said about the value of tests as supervisory instruments, as aids to school surveys, and as research tools, that frequently classroom teachers have not thought of them as of especial value to the everyday teacher in solving her own problems of the recitation. Yet this is where the most vital and important uses should be found. So this brief chapter will be given over to a consideration of some of these uses. They seem to fall into two general groups, dependent on the two great types of tests, namely, achievement and intelligence.

Uses of Achievement Tests

For purposes of comparison—Teachers frequently wonder just how their classes will compare in attainment with other classes in the same school system, or in the state, or in the United States. Before achievement tests were worked out, this knowledge was practically impossible. But now the giving of a standard test in penmanship or Latin or arithmetic makes possible a comparison with the norms which have been determined for the country in general for that particular subject, or part of the subject, so that one can easily see whether the

class is doing what should be expected of it at its particular point of advancement, if it is a typical class. In like manner, if norms have been determined for the state in which the work is done, or for the city or immediate community, a comparison which may be more valuable still can be made.

Comparisons with other classes within the system in which the teacher is working, within the same building, and even between different sections in charge of the same teacher, may also be made on a basis of the objective norms. Each of these comparisons has its own peculiar value in assisting the teacher to determine the relative attainment of any particular class at a given time. Where norms are not available except for the country at large, assistance in comparing special types of pupils may be frequently obtained by referring to various city, state, and county surveys, in which norms for large groups of children have been obtained. Also, by writing to bureaus of educational research in the larger cities in various parts of the country, norms may be obtained which will be of value for comparative purposes.

Another sort of comparison which is useful is that between the attainment of a class at the beginning and at the end of a semester's work, or at lesser intervals in the course of the semester. Such comparisons are decidedly stimulating to class work, if the interest of the class is aroused by having the purpose of the tests explained, and the goal of expected improvement set definitely before it. In such instances, the test itself is of course not set as the matter for drill, but is carefully kept from the pupils, to prevent any sort of unfair use of its content. This means especially that a pupil is

not to be allowed to retain a copy of the test at any time that it is given.

Alternative forms of nearly all tests are available for comparisons of the sort indicated, so that if there is any suspicion that pupils are "coaching" each other, or have kept copies of tests, the alternative tests may be given. They are constructed so that they will be of equal difficulty with the originals, and so will be entirely accurate for comparative purposes.

For diagnosis of teaching—The comparisons just suggested are valuable in themselves, by way of indicating whether classes are up to standard in their attainment. But it is even more important to be able to determine as a result of these comparative studies whether the teaching itself may not be improved. The result of the tests will show the effect of drill, for one thing. If the drill work has been ineffectual, this will be evident. If too much time has been given to such mechanical aspects, the undue proficiency of the class will show that there has been waste along this line.

When the class shows great strength along one direction, and weakness along another, the instruction will of course be directed to remedy the weakness, at the expense of less work along the lines where proficiency is evident. The comparative study will show whether the teaching has been too mechanical,—whether it has emphasized the drill aspect to the exclusion of the thought process, as when the progress of a class in arithmetic is tested for mechanics by the Courtis tests, and for thought by the Stone, or other reasoning test.

Methods of presentation may also be tested in this manner. Every teacher is looking for the best method of

instruction and the best way of measuring results of various methods is by the achievement test. Thus it will be clear that both strong and weak points in teaching may be diagnosed by the teacher herself, independent of a superintendent or supervisor, and then, of course, she will work out the necessary prescription to remedy any faults apparent.

For diagnosis of classes—Especially at the beginning of a term, the teacher wishes to know a good deal about the proficiency of her classes in their subject matter, and in general preparation for their work. She wants to know where their weak points lie, in order to direct her work where it will do the most good. She wants to know the sort of preparation her pupils have had in general for the work. For all of these purposes, various sorts of tests will be given.

Such a test as the Barr Diagnostic Test for American History combines a number of factors of this sort into one test, and so makes easier the determination of weak spots in the preparatory work of the class. But the teacher does not have to have a special diagnostic test to find out many of the things she needs. As reading is the basis for proficiency in all subjects, she can test the ability of the class in both oral and silent reading very easily. This information alone will indicate whether she can expect her class to interpret the printed page with facility, and so master their text books without great assistance, or whether she will have to make interpretation her major work. In like manner she can test for reasoning, for judgment, and for other general qualities, as well as for ability in specific subject matter.

Not only is this preliminary diagnosis of value, but also is it valuable and necessary to test advancement from time to time, by other than subjective tests. The test of advancement will reveal whether the class as a whole is moving together, or whether there are more or less well defined groups which seem to need special treatment, and which will justify the teacher in dividing them into sections for special instruction. Where the instructor has but one grade in a room, or, in high school, has several sections of the same subject, she will want to arrange her pupils so that groups of like proficiency can make equal advancement. The objective test is the best means for making this adjustment so that pupils can move forward in groups of like attainment.

Every set of classes in a school shows a good deal of what is known as "over-lapping." That is, when pupils in several grades are tested, it almost always is found that there are pupils in one grade who can attain the norms of a grade or grades higher, and frequently other pupils who seem to be unable to meet the norms of the grade in which they are placed. Thus in a spelling test, fourth grade children are found who can spell as well as the typical seventh, or even eighth grade pupil; and fourth grade pupils are also discovered who can not meet third grade norms. The tests will tell the teacher whether enough of a given class overlap the norms of the class above, in any subject, to make it possible, other things being equal, to advance this group into the work of the higher grade either in one, or in all subjects. A number of schools in the United States have been so organized that when these overlappings are found, rapid

moving or slow moving groups may be formed, to meet the situation. This of course means that promotion by grade is subordinated to promotion by subject.

For diagnosis of individual pupils—Closely connected with the use of the test for diagnosis of the class, is its use for determination of the difficulties and peculiarities of each pupil. While in general individual differences are not so marked as to preclude efficient class instruction, yet the more that is known about each child's weaknesses and strong points as well, the better success will the instructor have in handling the group. The test is to be studied especially in the light of each pupil's individual attainments, and points of difficulty. This study may be the means of clearing up entirely unsuspected troubles which otherwise would have continued to hamper the child, and to prevent proper advancement.

Especially valuable is this sort of individual diagnosis in the light of the various sorts of drill sheets and practice pads which have been designed in various subjects so that each pupil may drill upon his own difficulties, and do this independently of the class. The Courtis and Studebaker practice pads in arithmetic, and the Courtis practice exercises in handwriting, are examples of this sort of drill opportunity which carries each pupil at his own gait, without affecting the rest of the class.

The cases where pupils are found to overlap other classes to a marked extent, so that the treatment must be individual rather than group, call for special treatment. In such cases pupils who are far beyond the norm of the class in a special subject as arithmetic, may be placed with the advanced class in that subject for recitation work, keeping with the class in other subjects, but still

gaining time. And other pupils markedly deficient in some one subject, may be given extra drill by being placed in the lower class. Of course this entails an organization of the school to permit this sort of arrangement, but this is being made more and more commonly now, so that it is by no means an impossible, or even very difficult matter of administration, given the will on the part of the principal or superintendent to carry it out.

Many teachers feel that this use of the test to bring out the individual's special proficiency, or lack of it, is the most valuable use of the tests. Certainly it is an important one.

For setting standards for achievement—In Chapter II, it was pointed out that norms have been developed not only for the achievement to be expected of an average child in a grade, but also that norms have been worked out in connection with many tests as standards for achievement to be aimed at in connection with the work of a year or semester. This setting of a definite goal to be striven for by both the entire class, and superior pupils in the class is a great stimulus to both teacher and pupil. The interest and enthusiasm of the entire class can be aroused in the attempt to attain a standard which has been shown by the experience of other classes in the same or other communities to be within the bounds of probability for high grade work, as in no other way. To be sure the diagnosis of the class must have shown beforehand that these standards are reasonably to be expected under the right sort of teaching, for it is bad to work for a goal which seems, and for that very reason perhaps is, impossible of attainment for some particular group.

So the test may be used as a very potent means of stimulating the work of both the class and individual.

For pupil self-examination and rating—Many tests make possible the self-rating of the pupil. He can be taught to compare his work in composition or hand writing with the scales or tests posted on the wall of the school room, and to determine the value of his own specimens for himself, without intervention of the teacher. He thus can see his own weak points and can strive to remedy them in a way which is most valuable. Beating one's own record is an absorbing occupation in almost any sort of activity, and this idea can be capitalized in getting children to work against their own records in school subjects. When this self-rating is translated into a school mark, it goes far to reconcile the pupil with what he might otherwise think an unfair or prejudiced teacher's rating. And this is also reflected in the attitude of the parent. When the parent realizes that his child feels that his mark is obtained by his own comparison of actual work with a required goal, and is arrived at by objective methods, he is the more likely himself to be satisfied with it.

For promotion and marking—Tests may well be made the basis of promotions, using them judiciously. By this is meant the principle that no single test of any sort should be made the entire basis of promotion in school. There should always be a combination of factors, of which the test may be a very important one. In any subject, the work of the pupil from day to day is a very considerable factor, and should be estimated fairly, and, so far as possible, objectively, and taken as a very great element in the promotion of the pupil. Whether mere effort

should be taken as a basis is very questionable, but where there is doubt, effort of an unusual sort may well be the determining factor in deciding whether a pupil shall proceed to higher work.

In general, conduct is a factor which should not be considered in promotions, although it is sometimes felt, by a false pedagogy, to be a proper consideration. But after all things are considered, the result of a series of objective tests in a subject gives the most satisfactory basis for determining the real fitness of a pupil for advancement. Both pupil and parent can understand how the results of such tests are connected with the general norms to be expected of all children everywhere, and can see the failure or success of the child in meeting such a norm, where the subjective mark has no such visible method of explanation.

Uses of Intelligence Tests

For class diagnosis—After the teacher has given the achievement tests, and has found how her class stands in relation to the norms in any given subject, she is in danger of making some false assumptions relative to the class unless she has some further information. If her class falls below the expected norm, she may feel that her teaching has been a failure. This may be an unwarranted assumption, for the class may actually be of a low grade of mental capacity, and not able to make the norms expected of a class of average ability. In fact, her teaching may be above the ordinary, to have enabled the pupils to reach as high a standing as they have shown. Another teacher may plume herself upon a fine piece of teaching, when the very brilliancy of the group under

really strong instruction should have enabled it to reach a much higher rating than that actually received.[1] At once there becomes evident a need for some means of determining approximately the actual intellectual capacity of a class. The intelligence test has been devised to meet this need. We have already spoken of the various tests suitable for use in groups of pupils, and by giving two or three of these tests, the teacher can determine with some degree of accuracy whether any class is up to the normal expectation in ability to learn school work.

It is true that the technical difficulties both in giving and in interpreting intelligence tests are greater than for achievement tests. But the latest forms of these tests are being devised with the idea that they will be given by classroom teachers, and as will be shown in Chapter XVII, with proper precautions the teacher may feel that her conclusions are approximately correct. She may, then, by comparing the results of several of these group tests, make a diagnosis of the ability of the class which will go a long way in enabling her to determine what degree of proficiency she should reasonably expect of it in connection with the regular class exercises. This

[1] The significance of the intelligence factor in a class was recently brought out in a study designed to evaluate a certain, highly-advertised method of teaching pupils individually instead of in classes, with special emphasis upon freedom of choice of work by the pupil. The pupils were tested by standard tests in the various school subjects and found to be up to or above normal standard in all. But the children had also been given intelligence tests, individually. The average Intelligence Quotient for the class was found to be about 120! This denotes "superior intelligence," and means that the children averaged about two years beyond their actual ages, in intelligence. Similar tests were made in two successive years with similar results.

In the light of these facts, the children were accomplishing much less than should have been expected of such bright children, especially as the teachers were above the average in ability. The method of teaching, therefore, was not justified in the light of the results.

knowledge will enable her to plan her work more intelligently, and to prepare for overcoming difficulties which otherwise might have been unsuspected, or delayed in making their appearance, for here as in other affairs, "fore-warned is fore-armed."

Where this matter of intelligence testing is done by the supervisory force both time and trouble are saved to the teacher, for she has simply to call upon the proper authorities for the information which otherwise she would have to work out for herself. Strictly speaking, while achievement testing may be thought of as essentially a part of the work of the classroom teacher, intelligence testing is rather the work of the specialist.

For individual diagnosis—The intelligence test is especially valuable to the classroom teacher in assisting to solve the problem of the proper treatment of the child who is out of the ordinary. He may be unusually bright, or dull, or mischievous, or troublesome, or in some other particular atypical. The instructor wishes to know of this child whether his general ability is as indicated by his typical responses; whether the judgments of former teachers, and, perhaps, the child's own parents, are correct. For this, the intelligence test gives information not to be obtained in any other way. The result may indicate that the pupil has ability hitherto unsuspected; or that his supposed brightness is but a superficial pertness covering an actually dull intellect; or that the pupil's bad behavior comes from not keeping busy a really brilliant mind, so that his mental activity finds an illegitimate outlet in mischief and disorder.

Again, when given to an entire class, the intelligence test frequently uncovers a child of real brilliancy who

has been content to go with the group without in any way showing his real ability.[2] Such unsuspected "finds" would never have been known, had it not been for the intelligence test.

The overlapping described for the various classes in achievement occurs in the same way in ability. There will be children found in the fourth or fifth grades with the intelligence of normal eighth graders; and also pupils found in the fourth grade who in intelligence should be in the third or second.

Whenever individuals are discovered who are far in advance of their place in school, in most cases there should be a readjustment of work to their abilities, either by advancing them to a grade of work where their intelligence is given a real test, or by placing them in rapidly moving classes, so that they may make progress according to ability rather than by some fixed promotion period.[3]

[2] An interesting illustration of this fact recently came to my attention. Ann was a pupil in the fourth grade of a private school, this grade corresponding to her chronological age. She was transferred to the public schools and was given the Terman achievement tests and an intelligence test. Her intelligence quotient was found to be 161, and her achievement test scores were above the norms for the sixth grade! The girl was placed in the sixth grade of the public school with a short period of coaching on the omitted work of the grade missed. At the end of the second month in the sixth grade she is doing better than average work for that grade.

The opposite condition existed in the case of a boy in the same system who was one of several who were demoted one grade on account of low intelligence quotient for the grade involved. The parents of the lad objected strenuously to his demotion, but a few weeks later the boy himself bore this voluntary testimony to the principal: "Gee, I sure am glad I got sent back. You know I just could n't understand any of that eighth grade work. Now I am getting along real well."

[3] Frequent objection is made to a plan of segregation, on the ground that the placing of the slow pupils by themselves causes them to lose initiative, and removes the example of the brighter pupils. But this does not follow if the teachers are careful not to indicate that any

The dull pupil also will be better understood in the light of the intelligence test, for his difficulties will be diagnosed more particularly, and his strong points will very likely come out into relief, so that the result may be an entire re-direction of his instruction. In any case, the intelligence test will assist both in explaining difficult cases, and in revealing unsuspected strength, and perhaps, weakness, on the part of apparently normal, well-disposed pupils.

Of course, too much importance must not be attached to the intelligence test alone. This has been emphasized very definitely in the past few years, and the result has been that some persons have interpreted this protest against undue dependence upon such tests to mean that the tests were of no value. What is really meant is that the tests are not to supersede the results of classroom achievement, but to supplement them; that the tests are not to be used in place of all other information to be obtained about a child, but in the light of such information; that the child is not to be placed in school, or in other relations of life, simply on the basis of his I. Q. (Intelligence Quotient), but that his I. Q. is a means of interpreting otherwise undiscoverable factors of his personality.

The combination of the achievement and intelligence results—Accordingly, that there may not be too great

stigma is to attach to the slow class. In fact, the pupil finds that he is more at home in not being held up to comparison with the brighter children, and that he may well have a better chance to show real progress if he is competing with his own kind. Some one has said recently that it is always possible to get up a fat man's race, and the competitors take just as much interest as do the spectators. They are quite willing to compete in a race in which each feels that he has a fair chance to win. The parallel is a good one.

stress laid upon the child's proficiency in any one sort of test, the well-informed teacher of to-day attempts to rate the pupil in the light of his various attainments. By the intelligence test, she discovers his "mental age," that is, how his general ability to learn agrees with a norm worked out for presumably normal children of any given chronological age. Suppose that she finds that a child nine years and three months old, actually, tests on the intelligence scale as high as the "normal" child of ten years and six months, (actual age): This child is then said to have a "mental age" of ten years and six months. And his intelligence quotient is the ratio of his mental age to his chronological age, that is, the ratio of ten years, six months, to nine years, three months; or reducing to months, the ratio of 126 to 111. Expressed in a single figure, this ratio, 126 divided by 111, is 113.5, which is said to be the child's Intelligence Quotient. The child thus has an I. Q. of 114, we say in general terms, disregarding the fraction.

But the teacher has tested the child in other ways. She has found his ability in a Courtis Arithmetic Test, an Ayres Spelling Test, a Buckingham History Scale, and so on. She finds here also, his age in achievement in these subjects, as compared with his actual chronological age. And she works out ratios for these subjects, as between his real and his mental ages. Although usually the norms for achievement tests are given in terms of school grades, and not of years and months, yet by using the ages usually assigned as average ages for the school grades, a reasonably accurate index may be worked out. Thus, when a norm is given for an arithmetic test for

the seventh grade, it may be spoken of as the norm for the average age of a seventh grade child.

McCall, in his *How to Measure in Education,* has worked out a table based upon investigations by Ayres, Terman, and Kelley which gives the average age of a first grade child as 80 months, at entrance into school, and adds thirteen months to this figure for each succeeding year, as the investigators quoted seem to agree that the average time spent in each grade is from twelve to thirteen months. On this basis, the average age of a seventh grade child is 158 months, or if the norm is for the month of May instead of September, as usually happens, then the age would be 167 months. The achievement age of the child may then be compared with his chronological age, and the quotient resulting may be given a name—some psychologists call it the Educational Quotient, or E. Q. If this E. Q. is divided by the I. Q. of the individual, the result of this division, or the ratio of the E. Q. to the I. Q., has been called the Achievement Quotient, or the A. Q. This ratio may also be calculated by dividing a pupil's achievement age by his mental age. Franzen, McCall[4] and certain other workers have called it "Accomplishment Quotient." Thus A. Q. may be read either Achievement Quotient or Accomplishment Quotient.

Franzen has suggested that the Accomplishment Quotient be calculated for a combination of achievement tests instead of being based upon a single test such as one in reading. Buckingham and Monroe have devised a set of combined educational-mental tests arranged so

[4] *How to Measure in Education,* p. 85 ff.

as to facilitate the calculation of the A. Q. Pintner has also devised a similar combination of educational-mental tests, but provides for the calculation of a difference rather than a quotient. These tests give, of course, rather rough measures, but they go far toward solving the problem of diagnosis for a group of children, and so are suggested as measures worthy of trial.

CHAPTER IV

REQUISITES FOR GIVING OBJECTIVE TESTS

The earliest tests which were made public were so constructed that the typical classroom teacher would have had difficulty in using them. In fact, it was the early idea that only a trained psychologist should attempt to give or evaluate such measures. But as time has gone on, it has been found possible so to modify the nature of many of the tests, and to give such careful and adequate directions for their use, that almost any intelligent person with experience in handling children, can secure satisfactory results. In this chapter will be set forth some of the principal factors which enter into the use of tests, if they are to be accurate measures. After studying these various requisites, each teacher will have a very good idea of her own adaptability to the work.

There are three steps to be considered in testing: (a) the giving of the test; (b) the scoring of the papers; (c) the interpretation of the scores.

The giving of the tests—The requisites for giving tests properly come under four principal heads: (1) ability to follow directions; (2) personal poise and control; (3) pupil control; (4) proper environment.

(1) The person giving tests must be temperamentally fitted to follow directions to the letter. That teacher who is so fond of showing "individuality" that she can

never do things as others do them is temperamentally unfitted for testing. For if there be any deviation from the instructions which are to govern the giving of the test, there will be deviations in the results, and the scores will be worthless. Testing is scientific work, and must be done by scientific methods; one prime quality of scientific method is accuracy. This refers to the language in which test directions are given; to time limits which are imposed; to preliminary arrangement and preparation of pupils; to voice modulations and quality. The substitution of a single word in giving instructions or asking test questions may alter the entire idea desired. The inflection of the voice may give a pupil the hint to an answer or a cue to a situation or response in a way not at all intended. The addition of a quarter of a minute to the time allotment may have no appreciable effect, but it may change an entire class rating.

Therefore the teacher must follow directions to the letter. If unusual situations arise, she is to give the test as instructed, and in evaluating it, if there be doubt about the validity of the results, should submit them with the attendant circumstances to a trained expert for an opinion; or if there be no one of the sort available, to the author of the tests, or some college professor of educational psychology. Time elements must be observed to the second. The voice must be kept at as even a pitch as possible; voice inflections must not reveal any emotional disturbance in the teacher, which would in any way invalidate the intent of the directions.

All of this means a definite preparation for giving the test; a careful study of the directions; a rehearsal orally of the entire test before appearing before the class with

it. If possible, the teacher should practise by giving the test to other adults, and should also take it herself. Her preparation should result in an absolute familiarity with the test and with the possible contingencies which might arise in connection with it.

In like manner a study should be made of the best preparation, arrangement, and seating of the pupils who are to be tested. A very proper method is to arrange the group as for the proposed test, and then give the group a brief subjective test along somewhat similar lines to accustom them to the sort of thing which will come in the actual test, carefully refraining from anything which would be practice of the actual situations involved in the real test, except for the general arrangement and attitude of the children. Many tests are now constructed with preliminary exercises intended to give the children an idea of the sort of thing to be required of them, and used as preliminary to the main test to insure an absolute understanding of the requirement of the test. These are called "warming up" exercises, or "shock absorbers." Where these are available, they will be found to be of great assistance in preparation for the real test. Most tests are constructed with the thought that they will be given to regular class groups, so in many cases this preliminary arrangement is unnecessary. But the teacher should satisfy herself on that ground, so that she may not find an unfamiliar extraneous factor introduced which might have been foreseen and avoided.

(2) The teacher who is not self-controlled, who is lacking in poise, is likely to fail in giving the tests. Undue emotionalism of any sort is incompatible with scientific accuracy. The teacher who allows impulse to rule

her actions, who is overly sympathetic with the pupil who is having difficulty with his test, who can not resist giving hints, or "little helps," who is thinking more of the showing made by the pupil, than of the value of the test itself, is not to be trusted with such scientific devices. The teacher who yields to such tendencies can not make up for this laxity by making allowances in any other direction, without making a bad matter worse. A certain high school principal, very capable as an administrator, but without adequate knowledge of tests, a short time ago attempted to give a certain test, but decided that the time limit was too short, and so allowed double the time required. When told that this invalidated the test, he said: "I can fix that all right; I shall just divide the scores by two! !" A better example could not be given of failure to appreciate the scientific attitude. The teacher "who likes to do things her own way" and so modifies her manner and attitude as to defeat the purpose of the test, either by actually failing to follow directions, as has been already suggested, or by interpolating remarks which she thinks will be of assistance to the pupils, is also to be forbidden the use of the test. On the other hand, that teacher who in general has developed a feeling of antagonism or of fear in her classes, who is totally unsympathetic, harsh, or bad tempered, will find that the class will not respond properly to the test

The teacher who is even-tempered, self-controlled, master of herself in every way, will make the ideal examiner. Absolute honesty of purpose and attitude is of course presupposed. The type of person who "would

cheat in playing solitaire" is of course not fitted either to teach or to examine pupils.

(3) Test results mean nothing unless the class or group is under definite control. Order is essential to testing as to every successful school exercise. That teacher who does not know how to secure and maintain good order can not succeed in testing work. The group must be under such control that it is in sympathy with the purpose of the teacher, whether that be to give an ordinary recitation or to use an important test. Pupils work best in tests as in every other sort of work, when they are in as nearly a natural frame of mind, as little flurried or agitated, as possible. The teacher who can maintain sympathetic control will have such an attitude in the group, and will find that the result of the test comes as near being accurate and representative as can be desired.

Further, the examiner must be cautioned not to give a test when the group is agitated or "upset" by any untoward circumstances. Lapses from ordinary class conditions may lessen the control factor, and so tests should not be scheduled when there is likely to be any unusual condition affecting control. Thus, the day before a holiday, or the day of a crucial athletic contest, or the morning preceding the school picnic, or any such time, always lessens the control factor, and so should be avoided as a test day.

(4) This leads to the fourth factor, namely, the environmental conditions. The test can be best given when there is absolute quiet. If the ordinary classroom is so situated as to be affected by the noise from the

street or nearby industry, for purposes of testing an exchange should be made for the period with some other classroom better suited to the purpose. The change of environment of the strange classroom is not likely to be so serious a factor, except for very small children, as the presence of noise. The test should not be given when there is a recess for a part of the school, making a noisy playground situation which may inhibit the proper attention of the pupils to the test.

Interruptions must be avoided. So the test should not be held at a time when there is any likelihood of an interruption from visitors or from a fire drill, or when there is any possibility that the time ordinarily allotted may be curtailed, or danger of any other type of distraction which would invalidate results, either for the group or for individuals. The tests are so important that they should not be exposed to the danger of modification by any such preventable situations.

Any factors which might induce undue strain are to be avoided. Presence of a principal or superintendent may produce such a situation. In such cases, if it can be managed, these supervisory officers should be persuaded to remain out of the room. Pupils may be actually frightened by unusual stress being placed upon the importance of the test, and so unfitted for the best results. The less in general that is said to the groups about the importance of the tests, the less strain is likely to result. Certainly the attitude of some teachers in advertising that they are to give tests is not conducive to the most satisfactory results, or the best environmental influence.

The scoring of the papers—It will be remembered that

one of the great objections to the subjective type of test or examination is the variability of marks assigned to the papers by different teachers, or to the same paper by the same teacher on re-scoring. One of the great values of the objective test is that it tends to remove this difficulty. But this result will not hold, unless as much attention is paid to following directions in scoring, as in giving, the paper. The teacher will find it possible to obtain specific instructions as to methods of marking the tests, and must follow these instructions without deviation, or the result will be invalidated. A teacher may not altogether agree with the principle of marking used by the author of the test, but can not for this reason change the method of procedure. The standardization of a test means that absolutely uniform methods have been used both in giving and in scoring, and any departure from these methods will make the test worthless. For these reasons the instructor must remember that all elements of subjectivity must be eliminated.

One great advantage of most tests and scales is the ease with which they may be corrected. Not only are instructions given for the method to be followed, but devices are suggested for reducing the work to a mechanical system, in all cases where no element of judgment is involved, as in practically all of the intelligence tests, and in many of the achievement tests and scales. The drudgery of correction is therefore much less than in the older type of examinations, so much so that many teachers are modeling the regular subjective examinations upon the plan of the objective tests. For this reason, the teacher need not fear that the giving of the test will involve a great deal of extra time for correction.

After the tests are corrected, the next step is to tabulate the results. Many tests are accompanied by instructions for some method of setting down the marks of an entire group or class, in such a way as to be easily read and referred to. In general, such methods are the result of long experiment, and will enable the teacher to use her data to best advantage. In order to interpret the tests in the light of the plan of their construction and application, the suggested methods of tabulation and arrangement should be followed, even if they involve a somewhat different type of work than has been a part of the past experience of the teacher. This refers especially to methods of graphing curves or using other methods of charting results. Directions such as those accompanying the Courtis Practice Tests in Arithmetic, very clearly tell how to make such graphs, and give illustrations of them. Therefore the teacher will not find the task of making simple graphs a difficult one. In fact when she has once learned to read a graph intelligently, she will prefer this method of presenting results to any other.

The interpretation of the scores—After the papers are marked, and the scores tabulated, or graphed, the teacher is then ready to make use of them. Many persons think of this matter of interpretation as one involving a knowledge of abstruse statistics, profound psychology, and unusual experience and judgment. Truly enough, all of these factors are involved in making and standardizing the test. But the classroom teacher has had the way prepared by the scholar and the psychologist, so that she need anticipate no great difficulty in the more general and obvious forms of analyzing and applying the results. In the descriptions of various tests given in the follow-

ing pages, an attempt will be made to indicate the uses which may be made of these tests, and hints will be given as to conclusions which may be drawn from results. When the tests themselves are obtained, additional data will accompany them, and will further simplify the conclusions to be drawn.

A few terms commonly used in measurement and statistics must be mastered, in order to make the language of interpretation entirely intelligible. The median, quartile, deviation, probable error, and similar terms, are used so frequently that no teacher is justified in being ignorant of them. A brief statement of their meaning will be found in Chapter XVIII. More extended knowledge may be found from such treatises as Thorndike's *Mental and Social Measurements,* or Rugg's *Statistical Methods Applied to Education.*

In Chapter XVIII will also be found some illustrations of the way in which certain concrete results may be interpreted, which will give a better idea of certain forms of analysis than any description. Therefore, further discussion of this phase of testing will be left to the later chapters of the book.

The principal idea which this part of our treatise should leave in the mind of the reader is that giving and scoring of a test is worth nothing in itself. The use of the measure gives it its only value. Many persons have been quite content to say: "I gave the Buckingham test to my classes last year," and have felt that they should even be congratulated on their progressiveness, when they did not actually make use of one bit of data which might have been gained from the test. Therefore the teacher should look into the use which she is to make

of the measure, the way in which she is to interpret the results, familiarize herself with the real purposes of the tests, and then put all of this knowledge into practice, or she should not give them at all.

CHAPTER V

SPELLING

The beginning of work in educational measurement was made by Dr. J. M. Rice in 1897. His report of an investigation in spelling at a meeting of the National Education Association in St. Louis marks the first public announcement of the modern idea of measuring the results of teaching. Rice selected a list of fifty words and submitted them to a number of schools in order to determine the effect of different amounts of drill on ability to spell. His list of words was not scientifically selected nor was it standardized in the sense that he knew what score an average third or fifth grade child should make. Yet it is worthy of notice that Rice used this first objective test as a means for determining the relative effectiveness of different methods of teaching. Instead of trying to settle the problem of methods by the traditional plan of opinion and debate Rice offered a measuring instrument for the exact determination of efficiency. This idea was new to teachers and administrators and it opened the way to a careful scientific study of the problems involved in the learning and teaching of spelling. Let us further consider what some of these problems are.

Spelling depends upon the formation of certain definite associations as the result of experience. In spelling a word we recall how the word looked or sounded when

we spelled it, or when someone else spelled it for us; what letters the word contained and their order; or more likely motor imagery carries the hand along in the writing process without much thought about the spelling. This holds true for the average adult. With the child the problem of the formation of the correct associations for spelling is a difficult one. In the English language there are many sounds for some of the letters of the alphabet and some sounds are represented by different letters. Silent letters further complicate the spelling process. For this reason the child must form the correct association for the spelling of each word he uses. Most of his spelling is used in written work, hence the prevalence of motor control in spelling. These are some of the problems the pupil will meet in learning to spell.

The different methods for teaching spelling may be classified under two general headings. It may be taught in connection with reading, composition and the other school subjects by what is called the incidental method. On the other hand it may be divorced from the other school subjects and taught by the drill method. The words may be spelled orally or written. They may be used in sentences or out of context. A further problem of importance to the teacher is the selection of words that should be taught. There are more than 350,000 words listed in the New International Dictionary but Jones [1] found only 4,532 different words used by school children in their compositions. Terman [2] gives the number of words that the average eight year old child can

[1] See page 66.

[2] L. M. Terman, *The Measurement of Intelligence*, p. 226, Houghton Mifflin Co.

define as 3,600. Just what words, therefore, are to be selected by the teacher for spelling drill? These are some of the problems the teacher must solve in order to teach spelling efficiently.

These problems are not theoretical but practical problems of the classroom teacher. Stated more concretely, the teacher wants to know whether she is using the best methods in teaching spelling, whether she is devoting too much or too little time to spelling or whether she should teach spelling along with the reading, or, more important still, whether she is teaching the children to spell the words they most need to know. She may be most positive in her own opinion and give no further consideration to the problem. Unfortunately the strength of her opinion may bear no close correlation with the facts in the case. If the teacher attempts an answer to the problem by testing her pupils by the traditional methods, she will select a list of words from the speller, the reader, or elsewhere and give them to the class to spell. One child may make 90, another 65, and another 40. But what should they make? That depends upon the words. The only way to know what the pupil should make on a spelling test is to use a standard test in which we know what score a child of a certain age or grade should make. Some of the more important of these standard tests will now be described.

The Ayres Spelling Scale

Description of the Scale—This spelling scale was devised by Leonard P. Ayres of the Russell Sage Foundation. It consists of 1,000 words arranged in 26 columns with from two to eighty-two words in each column. The

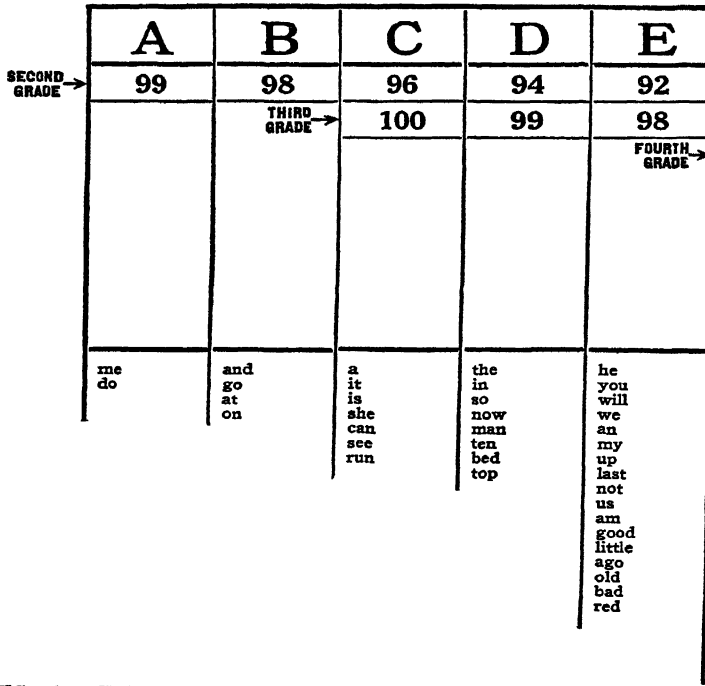


FIG. 1. SECTION OF THE BUCKINGHAM EXTENSION OF THE AYRES SPELLING SCALE

All the words in each column are of approximately equal spelling difficulty. The steps in spelling difficulty from each column to the next are approximately equal steps. The numbers at the top indicate about what per cent of correct spellings may be expected among the children of the different grades. For example, if 20 words from column H are given as a spelling test it may be expected that the average score for an entire second grade spelling them will be about 79 per cent. For a third grade it should be about 92 per cent, for a fourth grade about 98 per cent, and for a fifth grade about 100 per cent.

The limits of the groups are as follows: 50 means from 46 through 54 per cent; 58 means from 55 through 62 per cent; 66 means from 63 through 69 per cent; 73 means from 70 through 76 per cent; 79 means from 77 through 81 per cent; 84 means from 82 through 86 per cent; 88 means from 87 through 90 per cent; 92 means from 91 through 93 per cent; 94 means 94 and 95 per cent; 96 means 96 and 97 per cent; while 98, 99 and 100 per cent are separate groups. Similar meanings attach to the limits of groups below 50 per cent. Thus 42 means from 38 through 45 per cent; 34 means from 31 through 37 per cent; 27 means from 24 through 30 per cent; 21 means from 19 through 23 per cent; 16 means from 14 through 18 per cent; 12 means from 10 through 13 per cent; 8 means from 7 through 9 per cent; 6 means 5 and 6 per cent; 4 means 3 and 4 per cent; while 2, 1 and 0 per cent are separate groups.

By means of these groupings a child's spelling ability may be located in terms of grades. Thus if a child were given a 20 word spelling test from the words of column O and spelled 15 words, or 75 per cent of them, correctly, it would be proper to say that he showed fourth grade spelling ability. If he spelled correctly 17 words, or 85 per cent, he would show fifth grade ability, and so on.

AB	AC	AD	AE	AF	
← THIRD GRADE					
2	1	0	← FOURTH GRADE		
6	4	2	1	0	← FIFTH GRADE
12	8	6	4	2	← SIXTH GRADE
21	16	12	8	6	← SEVENTH GRADE
34	27	21	16	12	← EIGHTH GRADE
50	42	34	27	21	← NINTH GRADE
combustible *guarantee* *incessant* *lieutenant* *occurrence* *pneumonia* *proficiency* *villain*	*abyss* *cantaloupe* *embarrassment* *poultice* *sovereign* *syndicate*	*appendicitis* *chauffeur* *hippopotamus* *maneuver* *miscellaneous* *penitentiary* *souvenir*	*hallelujahs* *inflammable* *rhinoceros*	*conscientious* *discernible* *dissension* *jardiniere* *naphtha* *rendezvous*	

FIG. 2.

This scale is not a test but a list of words from which a teacher can make a test. The words in any column are approximately equal in difficulty; and it is best, therefore, to choose all of the words for a test from a single column.

Twenty words are enough to secure a reasonably reliable measure of the spelling ability of a class; but for such a measure of the ability of an individual 50 to 100 words will be required. Thus, owing to the fewness of the more difficult words, it may be necessary in testing upper grades to use words from more than one column. In such cases the differences in difficulty must be recognized.

In order that the words may be difficult enough really to measure spelling ability, they should be selected from columns for which the standard per cent of correct spellings is close to 50—say between 50 and 66.

The most appropriate measure of spelling ability is secured when the words are dictated in sentences at approximately the standard rate of handwriting for the grade in question, no test word occurring at the end of a sentence. The placement of words on this scale, however, is on the basis of returns from column dictation. Children spell more accurately when they write words in columns than they do when they write them in sentences. If, therefore, words are dictated in sentences, as suggested, results may be expected to be somewhat lower than the scale indicates. It was found that the words in each column from A through G fell three columns to the right when dictated in sentences (untimed); that those in columns H through Q fell two columns to the right; and that those in columns R through V fell one column to the right. No difference, due to dictation in sentences rather than in columns, appeared to exist for words harder than those in column V.

The 505 words added to the Ayres Scale by Buckingham are printed in italics. They were not chosen, as Ayres' words were, according to frequency in use in written discourse, but rather according to agreements among spelling books. They are not, therefore, offered as constituting a fundamental vocabulary in the same sense as do the original 1,000 words selected by Ayres. The original words of the Ayres Scale are printed in Roman.

Copies of this scale may be obtained from the Bureau of Educational Research, University of Illinois, Urbana, Ill.

words in each column are of approximately equal difficulty and the columns are arranged in order of difficulty with the easiest words in the first columns. The larger number of words are in the middle columns and there are a smaller number of words in the columns toward either end of the scale.

Derivation of the Scale—The first problem in the construction of a spelling scale consists in the selection of the words to be used. Ayres began by listing 368,000 words found in business letters, newspapers and literature. From this list he selected the 1,000 words recurring most frequently—in these three sources—as representing the most commonly used words in the English language. It is interesting to note in this connection that fifty different words appeared so frequently that they made up about half of the total list of material examined.

The next problem was the arrangement of these 1,000 words into a scale. In order to determine the relative difficulty of the different words they were submitted to 70,000 children from the second to the eighth grades in representative schools in widely separated parts of the country. On the basis of the number of times a word was misspelled the words were arranged into twenty-six lists. The words most often spelled correctly were placed in column "A" at the beginning of the scale and the most frequently misspelled words in column "Z" at the end of the scale.

Method of Using the Scale—The teacher in using the scale selects a list of words from one of the Ayres columns. Any number of words may be selected, but if reliable scores for the individual members of the class are desired at least twenty words should be chosen. In general, it

is best to select words from a column in which about seventy-five per cent of the spellings are expected to be correct. This gives sufficient range for both the best and the poorest spellers in the class. The words may be given either in or out of context. The pupils' papers are scored in the usual way by determining the per cent of words spelled correctly.

Norms [3] are given at the top of the scale for each list of words. At the lower end of the scale norms are given only for the lower school grades. In the middle portion of the scale norms for as many as four or five grades are given for each list of words. At the upper end of the scale norms are given only for the upper grades.

Buckingham Extension of the Ayres Scale

Description and Derivation of the Buckingham Extension—Professor B. R. Buckingham has made an extension of the Ayres list of 1,000 words by the addition of 505 additional words. These words were derived from the relative frequency of occurrence in a number of spelling books. They are, therefore, as Buckingham points out, not strictly an extension of the child's fundamental vocabulary. The difficulty of these words was determined and the words placed at the end of the Ayres columns. In general the additional words occur in the middle and upper end of the new scale.

[3] By norm is meant the score to be expected from a child in any given grade. For example: the norm for the sixth grade on the Ayres Scale for the words in Column "S" is 73. That means that the normal sixth grade child should make a grade of 73 in spelling a list of words taken from this column. These norms were derived by the number of correct spellings for each word by the 70,000 children for the different school grades as described above.

Function[4] *of the Scales*—The Ayres and the Buckingham-Ayres Scales are more than ordinary measuring devices. The method of selection of the material makes it fundamental teaching material. In other words the classroom teacher can well afford to drill her pupils on this list of 1,505 words instead of the usual large number of less frequently used words occurring in the ordinary spelling book. As Ayres states, the children should be so thoroughly drilled on the words that the scale would no longer be a measure of spelling ability.

While the list of words, because of the subject matter from which the words were chosen, may not be representative of the written vocabulary of children[5] it is much more so than usual spelling lists. The scale has a further advantage in its simplicity in giving and scoring. It is one of the relatively few standard tests that the average grade teacher can administer without special training in technique of giving tests.

Monroe's Timed Spelling Tests

Description of the Tests—Monroe selected his words for his test from the Ayres spelling list and placed the words in sentences. There are three tests. Test I is composed of 22 sentences for use with the third grade and 22 sentences for use with the fourth grade. Each group of sentences contains fifty words from column "M" of the Ayres list. Test II is similar, the words being taken from column "Q" of the Ayres list. One set of sentences

[4] The authors appreciate that this is a very free use of the word "function." It is here used in the sense of the evaluation and criticism of the tests and scales.

[5] For a more representative list see reference to Thorndike's *The Teachers Word Book* in a later part of this chapter.

is for use with the fifth grade and the other set for use with the sixth grade. Test III contains words from columns "S," "T" and "U" of the Ayres list imbedded in two groups of sentences, one for use with the seventh grade and the other for the eighth grade.

Method of Giving the Tests—These tests are timed tests in that the sentences are to be dictated by the tester at a certain rate. This rate is ten per cent slower than the average of handwriting as determined by Freeman for each of the school grades. The sentences must be read very distinctly with no repetitions. If a child can not complete the sentences within the time, he is to write as much as he can and then go to the next sentence. He is told before the test begins that if there are any words that he can not spell he is to omit them.

Scoring the Papers—The words taken from the Ayres list are italicized in the Monroe tests and only the italicized words are counted in scoring the papers. Since there are fifty words in each list, two points credit is given for each word spelled correctly. Norms for the different grades are given as follows:

NORMS FOR MONROE'S TIMED SPELLING TESTS

School Grade	3	4	5	6	7	8	9	10	11	12
Monroe Norm	56	78	66	80	70	84	86	90	94	96

Thus a third grade child should make a score of 56 on the test for the third grade and a seventh grade child should make 70 in the test for the seventh grade.

Function of the Tests—The tests have the advantage that the material is presented in context which is the most natural way of spelling. Very seldom is a person called upon to spell a word except in writing. For this reason these tests are superior to most of the other

spelling tests in this respect and the teacher wishing to determine the ability of her class to spell words in sentences should use the Monroe tests. The tests may be criticized on the basis of the timing of the rate of dictation. Although the rate of dictation is slightly slower than the average writing rate of children as determined by tests of 6,000 children in each of the school grades, it is too fast for many children. Ayres[6] has shown the great overlapping in speed of handwriting within any school grade. For example, in the fifth grade some children write twenty times faster than other children in the same grade. Approximately thirty-one per cent of the eighth grade children write no more rapidly than the average fifth grade child. Monroe says that this is not a serious fault of the scale since the words from the Ayres list are placed in the earlier parts of the sentences. This is only partially true. It is certain that many slow writers will make low scores on these tests not because of poor spelling ability, but because of slow writing. Whenever possible any factor to be measured should be isolated from other complicating factors. In these tests spelling ability is complicated by the introduction of the factor of speed of handwriting.

Sample Sentences from Monroe's Timed Spelling Test for the Fifth Grade

SECONDS	
60	The *president* gave *important information* to the men
48	The *women* were present at the time
19	The *entire* region was burned over
49	The *gentlemen declare* the *result* was printed
30	*Suppose* a *special attempt* is made

[6] See the Ayres Handwriting Scale, Russell Sage Foundation, New York City.

Iowa Spelling Scale

Derivation and Description—This test was devised by Ernest J. Ashbaugh especially for the measurement of the spelling ability of elementary school pupils of Iowa.[7] The scale includes 2,977 words taken from the written correspondence of Iowa people. The difficulty of the different words was determined by a total of nearly 4,750,000 spellings by children from each of the school grades. The words are arranged into twenty-five groups, each group varying from the one just above or below by approximately equal differences in difficulty. The scale is divided into three parts: part one for use in the second, third and fourth grades; part two for use in the fourth, fifth and sixth grades; and part three for the sixth, seventh and eighth grades.

Methods of Using the Scale—The author suggests that the teacher may use the scale in any one of three ways: (1) It may be used as a minimal list of words which the children of the elementary grades should be taught. (2) It may be used as an instrument for measuring the comparative skill with which children can spell a certain list of words as compared with the average for the State of Iowa. Norms for each grade are given above each list of words. Used for this purpose it is a real spelling scale. (3) It may be used as a measure in teaching. By comparing scores from time to time a teacher can measure her success in teaching spelling.

Function of the Scale—This spelling scale presents a very practical means for providing the classroom teacher with both a measuring scale and subject matter in spell-

[7] University of Iowa Extension Bulletin, No. 53, 54, 55, Iowa City, Iowa.

ing. The number of words is large enough to include practically all the common words in a child's vocabulary. The method for using the scale is the same as that ordinarily used by the teacher in written spelling. The teacher needs only to compare the scores for a pupil or a class with the norm to determine whether the pupil or class is above or below normal in spelling ability.

Material of English Spelling—Jones

Dr. W. F. Jones made a study similar to that made by Ayres except that the sources of his material were compositions written in school. He listed the words used by 1,050 children; approximately 150 from each grade; from four schools in widely separated parts of the United States. In all about 15,000,000 words were listed. But this list represents only 4,532 different words. These words are arranged in the Jones list by grades by listing each word in the lowest grade in which at least two per cent of the pupils used it.

From this list Jones selected the 100 words most often misspelled and arranged them into a list called the "One Hundred Spelling Demons of the English Language." Children should receive special drill in the correct spelling of this list of common words.

which	can't	guess	they
their	cure	says	half
there	loose	having	break
separate	lose	just	buy
don't	Wednesday	doctor	again
meant	country	whether	very
business	February	believe	none
many	know	knew	week
friend	could	laid	often
some	seems	tear	whole

been	Tuesday	choose	won't
since	wear	tired	cough
used	answer	grammar	piece
always	two	minute	raise
where	too	any	ache
women	ready	much	read
done	forty	beginning	said
hear	hour	blue	hoarse
here	trouble	though	shoes
write	among	coming	to-night
writing	busy	early	wrote
heard	built	instead	enough
does	color	easy	truly
once	making	through	sugar
would	dear	every	straight

The Teachers Word Book

The Teachers Word Book was compiled by Dr. Edward L. Thorndike to represent a more complete and satisfactory list of the most commonly used words in the English language. It consists of [8] "an alphabetical list of 10,000 words which are found to occur most often in a count of (1) about 525,000 words taken from the literature for children, (2) about 3,000,000 words from the Bible and English classics, (3) about 300,000 words from elementary school text books, (4) about 50,000 words from books about cooking, sewing, farming, the trades and the like, (5) about 90,000 words from daily newspapers and (6) about 500,000 words from correspondence. In all forty different sources were used.

The words are listed alphabetically and their relati frequency indicated by numbers at the side. The 1,000 words used most frequently have a credit number of 49 or more uses.

[8] See the introduction to *The Teachers Word Book*, E. L. Thorndike, Bureau of Publications, Teachers College, Columbia University, New York City.

The values of such a list of words as set forth by Thorndike are: (1) To inform the teacher what words in the reading lesson should receive the most attention. Many words are found in the readers that are not used frequently enough to warrant special study. The Word Book gives the teacher a method for selecting the words in the lesson for careful study. (2) It may be used by the less experienced teacher to provide her with that knowledge, both of the importance of words and of their difficulty, which the expert teacher has acquired by years of experience with pupils and books. (3) The Word Book may be used as a convenient place to record any useful facts about the words contained therein. (4) It may be made the basis for the construction of spelling lists of the most common English words. In fact it is the most carefully selected and complete list of its kind in the English language.

Materials Needed

1. The Ayres Spelling Scale for grades 2 to 8. Only one copy of scale needed. Price 5 cents. Russell Sage Foundation, New York City.
2. The Buckingham Extension of the Ayres Spelling Scale for grades 2 to 8. Only one copy needed. Price 14 cents. Public School Publishing Co., Bloomington, Ill., or *Spelling Ability, Its Measurement and Distribution*, B. R. Buckingham, Bureau of Publications, Teachers College, Columbia University, New York City.
3. The Monroe Timed Spelling Tests for grades 3 to 8. Test No. 1 for grades 3 to 4, test No. 2 for grades 5 to 6, test No. 3 for grades 7 to 8 and high school. Only one copy of the test needed for use in the grades indicated. Price 4 cents per test or single set 12 cents. The Public School Publishing Co., Bloomington, Ill.
4. Iowa Spelling Scale (Ashbaugh) for grades 2 to 8. Only one each of the three scales needed. See University of Iowa Extension Bulletins, No. 53, 54 and 55.

5. The Jones Spelling Scale for grades 2 to 8. Only one copy necessary. See *Concrete Investigation of the Material of English Spelling* by N. F. Jones, University of South Dakota, Vermillion, S. D., 1913.
6. *The Teachers Word Book* (Thorndike) Bureau of Publications, Teachers College, Columbia University, 1921.

Selected References

Ayres, L. P., *A Measuring Scale for Ability in Spelling,* Russell Sage Foundation.

Cornman, O. P., *Spelling in the Elementary School,* Ginn & Co., 1902.

Hollingsworth, L. S., *The Psychology of Special Disability in Spelling,* Bureau of Publications, Teachers College, Columbia University, 1918.

Rice, J. M., "The Futility of the Spelling Grind," *The Forum* (1897), pp. 163 and 409.

Sixteen Spelling Scales Standardized for Use in Secondary Schools by Earle Hudelson and others, Teachers College Record, Vol. 21, p. 337 ff., 1920.

Wallin, J. E. W., *Spelling Efficiency in Relation to Age, Grade, Sex, and the Question of Transfer,* Warwick & York, Baltimore, Maryland, 1911.

CHAPTER VI

HANDWRITING

Handwriting like spelling is largely a drill subject. It differs from spelling in that the product varies qualitatively rather than quantitatively. This does not mean that the quality of handwriting is not objective and can not be measured. The great difficulty in the measurement of handwriting has been the construction of an objective unit of measure. Teachers have not agreed on what constituted good handwriting. More serious than this is the fact that teachers have little conception of what constitutes a difference of, for example, 10% in samples of handwriting. Different teachers in grading the same sample of handwriting might vary even 50% or more in the grade given the sample; or the same teacher in re-grading the same sample might easily vary greatly in her marking. The writer recalls an interesting illustration. As county school examiner a set of papers in handwriting was being scored by the traditional method. During the grading a recess of an hour and a half was taken and by mistake the papers already graded were re-graded and the grades listed on a second score sheet. On comparing the two sets of grades for the same papers it was found that these prospective teachers profited by the examiner's hour and a half rest by receiving on the average more than 10%

higher scores in the second grading. This is only one illustration of a commonly known fact that the teacher or supervisor has no definite standards by which samples of handwriting may be compared.

In the construction of a handwriting scale the first problem is the determination of what constitutes successive stages of quality. In other words, just what is a quality of 50 or 75? If quality depends upon the style of the handwriting it may be necessary to present samples of different styles in the scale. Measurement in handwriting consists in the comparison of a given sample with a known standard or set of standards. A sample to be measured may be placed on a scale consisting of samples of handwriting with quality scores ranging from 20 to 90 and by comparison the sample is given the score of the quality which it most nearly resembles on the scale. By the use of a scale the teacher transfers her hazy and variable notions of quality values into an objective reality.

The average classroom teacher tends to make too little difference between the grades of her pupils. This is well shown by Ayres[1] in a study of the ratings given applicants by Civil Service examiners. By re-scoring these papers by the use of a scale it was found that while the examiners had ranged the grades from 60 to 95, the papers really varied from 20 to 90. The table on p. 72 gives a comparison of the two sets of grades.

Another matter arising in connection with the problem of quality is that of rate of writing. These problems are interdependent and both have received consideration

[1] *A Scale for Measuring the Quality of Handwriting of Adults*, Russell Sage Foundation, New York City.

Quality as rated by Civil Service Examiners.	*Quality as measured on the Ayres Scale—based on legibility.*
60	20
65	30
70	40
75	50
80	60
85	70
90	80
95	90

in the construction of most of the handwriting scales. By reference to these scales the teacher may find out whether or not she is placing the correct relative emphasis on speed and quality.

Enough has already been said to indicate many of the uses of a standardized handwriting scale to a classroom teacher, but there are still other ways in which she may use scales. Only by comparison of her class with the norms of the scale can the teacher know whether her pupils are up to or above standard in handwriting. In case they are below standard, by the use of a diagnostic scale she may discover in what specific respects they are poor and concentrate drill on these. In case the pupils are above normal the teacher may well spend less time on handwriting and devote more time to other subjects.

In connection with the problem of standards, Koos[2] found that four-fifths of 826 judges considered 60 on the Ayres scale adequate for the average person and even for most vocations, in handwriting. In fact he says that there is considerable justification for setting the ultimate standard for most persons as low as 50. No doubt a

[2] L. V. Koos, "The Determination of Ultimate Standards of Quality in Handwriting in the Public Schools," *Elementary School Journal*, Feb. 1918, p. 422.

higher score should be demanded of bookkeepers, bank clerks, elementary teachers and some others, but hardly in excess of quality of 70 of the scale should be demanded of anyone.

One caution should be given to the teacher in her first use of a handwriting scale. She may find almost as great variation in two markings of the same papers by herself, or between her markings and the markings of the same papers by another teacher, with the use of the scale as without. But a short period of practice in the use of the scale, especially by a group of teachers discussing the points of variability, will soon produce remarkably little variability in the assignment of grades to specimens of handwriting by a teacher and little variation between the grades of different teachers. Professor C. T. Gray [3] reports an experiment in which three students who had no experience in teaching or the use of scales were given practice in grading samples of handwriting by the use of the Ayres scale. Twenty-five samples were graded each week by each person. After the grading, conferences were held to compare grades and discuss difficulties in grading. The average variation between the highest and the lowest grades for the twenty-five samples for the first week was 20.4. This was considerable variation but probably not more than would have been found in grading without a scale. By the fifth week the variation was reduced to 12.7, and for the fifteenth week it was only 3.6. It seems from this experiment that a person without experience, by grading three or four

[3] C. T. Gray, "The Training of Judgment in the Use of the Ayres Scale for Handwriting," *Journal of Educational Psychology* (1915) Vol. 6, pp. 85-95.

hundred specimens of handwriting, could reduce the variability to almost a negligible factor. No doubt the average teacher accustomed to grading could do as well in a shorter time. If this be true, the results would highly justify the small amount of effort required.

With this general discussion of handwriting scales let us pass to a study of some of the more important scales.

The Freeman Analytical Handwriting Scale

Description of the Scale—The scale consists of five separate parts, one for measuring each of five elements of handwriting. The elements are (1) uniformity of slant, (2) uniformity of alignment, (3) quality of line, (4) letter formation and (5) kinds of spacing. The scale consists of three samples representing different degrees of quality in each of the five elements. The poorest sample in each case is given a grade of 1, the medium sample a grade of 3 and the best sample a grade of 5.

Method of Scoring—Any sample of handwriting to be graded is scored on each of the five elements of the scale separately and independently of the others. The sample is first given a grade between 1 and 5 on uniformity of slant by comparing it with the three samples for this part of the scale. In this scoring all factors of the handwriting other than the uniformity of slant are to be entirely neglected by the grader. After this scoring is completed the paper may be graded in each of the other four elements of the scale in turn. The final score for any paper is the sum of the separate scores on each of the five elements except that the score for (4) letter formation, is doubled. For example, a paper may be graded as follows: Uniformity of slant, 4; letter forma-

tion, 4; uniformity of alignment, 3; kind of spacing, 2; quality of line, 3. The final score for the specimen would be 4+(2×4=)8+3+2+3=20. The scale may be used for scoring a single paper or for a class. When a class is being scored it is better to grade all the papers on one of the five elements before passing on to the scoring of the next element.

Norms—Not only quality but speed of handwriting should be measured. Freeman gives the scores to be expected of children in grades 2 to 8 in both quality and speed of handwriting. These norms are as follows:

NORMS FOR FREEMAN HANDWRITING SCALE

Grade	2	3	4	5	6	7	8
Quality Score[4]	17.9	18.4	19.0	20.0	20.8	22.0	23.0
Speed (letters written per minute)[5]	36	48	56	65	72	80	90

Function of the Scale—In many respects the Freeman Handwriting Scale is one of our most ideal measuring scales. The characteristics of good or poor handwriting have been carefully analyzed. Each character is independently measurable by reference to given standards. This makes the scale highly objective. Definite norms have been determined for both speed and quality, but more important still, for the classroom teacher is the fact that the scale is diagnostic. In other words, the teacher by using this scale may find out not only whether a pupil or grade is up to or above standard but also in what elements or qualities the pupil or pupils are superior and in what ones they fail. It is very important for a

[4] *The Teaching of Handwriting*, Houghton Mifflin Co., Chap. V.
[5] Fourteenth Yearbook, National Society for Study of Education.

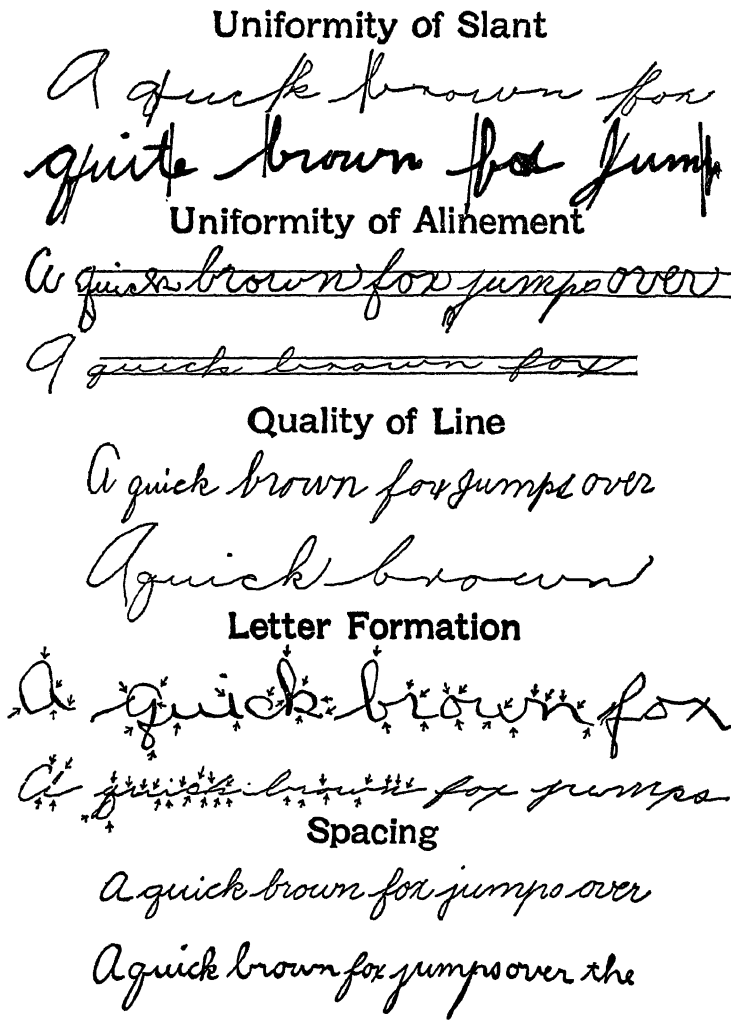


FIG. 3. SHOWING SAMPLE FOR VALUE 1 FROM THE FREEMAN ANALYTICAL HANDWRITING SCALE *

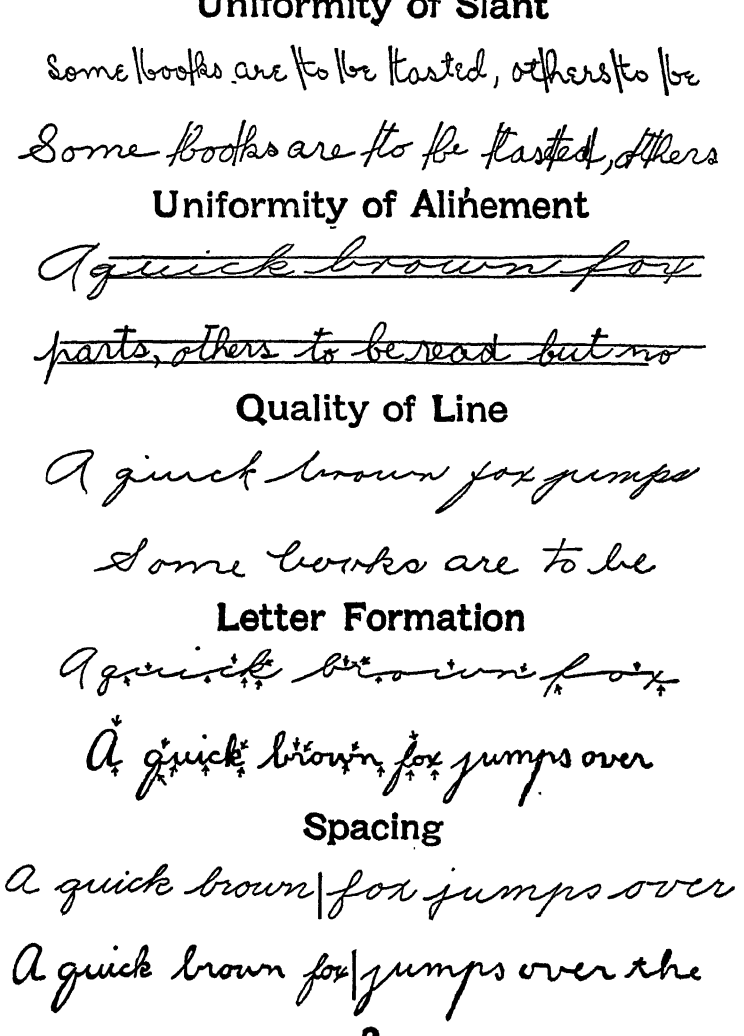


FIG. 4. SHOWING SAMPLE FOR VALUE 3 FROM THE FREEMAN ANALYTICAL HANDWRITING SCALE

teacher to know that a child's handwriting is good in quality of letters but poor in alignment and letter spacing. The teacher knowing these facts may immediately set about remedying the defects. Indeed these facts may easily be pointed out to the child to his very great advantage. By noting defects and by comparison with good copies the pupil can go a long way in his own improvement. To the teacher is left the problem of directing the pupil in the most efficient method of making these specific improvements. Furthermore, the scale is diagnostic in the sense that due relation between speed and quality is indicated. The pupil may be producing beautiful specimens of handwriting but at the expense of speed. This is inefficient as a general procedure whether deliberate or accidental.

The Freeman Scale has not been used as widely as some of the other handwriting scales. This is largely due to the fact that it has not been used to diagnose faults in handwriting but more often as a measure of general merit in handwriting. There are other scales that do this better and more easily than the Freeman Scale.

The Ayres Measuring Scale for Handwriting

Description of the Scale—The latest form of the Ayres Scale, called the Gettysburg Edition, consists of a series of eight samples of handwriting varying by increments of ten from the poorest sample with a value of 20 up to the best with a value of 90. Each sample consists of the first few lines of Lincoln's Gettysburg Address.

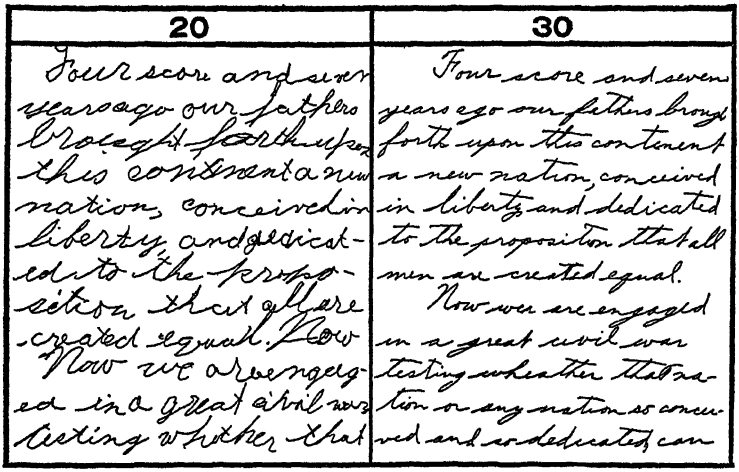
Derivation of the Scale—The Gettysburg Edition of the Ayres Scale is the outgrowth of two earlier scales.

The first of these was known as the Three Slant Edition. It was devised in 1912 as a result of the study of about 1,500 samples of children's handwriting taken from representative schools in 38 states. These samples were ranked by 10 investigators on the basis of the time required to read the selection. That is, the quality of each sample of handwriting was determined by the degree of legibility as shown by the rate of reading by these ten judges.

Three sets of samples were then selected and assigned equivalent values of 20, 30, 40, 50, 60, 70, 80 and 90. One set was written in ordinary slant, another set in vertical writing, and the other in backhand. The samples in each set represent qualities such that each one is as much better than the one that precedes it as this one is better than the one that precedes it. In other words the units of the scale are equal and the difference between a score of 20 and 30 is the same as the difference between a score of 80 and 90. While this may seem very simple and easy it is not. Ayres accomplished this according to the distribution of scores on the normal curve of frequency. The three slant scale had certain advantages but was discarded in favor of a one slant scale.

The Gettysburg Edition is similar to the three slant scale in nature and construction except that there is only one sample of each quality, written in ordinary style of handwriting with the usual amount of slant. The material for each quality of the scale was the same. The material as already stated was taken from the first few lines of Lincoln's Gettysburg Address.

Method of Using the Scale—The children to be graded in handwriting are first drilled on the first three sentences



This scale for measuring the quality of handwriting is a revised edition of a scale first published in 1912 and subsequently reprinted 12 times with several minor revisions and with a total of 62,000 copies. The purpose of the changes introduced in the present edition is to increase the reliability of measurements of handwriting through standardizing methods of securing and scoring samples, and through making numerous improvements in the scale itself designed to reduce variability in the results secured through its use. The present scale may be referred to as the "Gettysburg Edition" in order to distinguish it from other editions. The original or "Three Slant Edition" and the scale for adult handwriting are not superseded by the present scale. Copies of any of the three scales may be secured for five cents each, postpaid.

To secure samples of handwriting the teacher should write on the board the first three sentences of Lincoln's Gettysburg Address and have the pupils read and copy until familiar with it. They should then copy it, beginning at a given signal and writing for precisely two minutes. They should write in ink on ruled paper. The copy with the count of the letters is as follows:

Four 4 score 9 and 12 seven 17 years 22 ago 25 our 28 fathers 35 brought 42 forth 47 upon 51 this 55 continent 64 a 65 new 68 nation 74 conceived 83 in 85 liberty 92 and 95 dedicated 104 to 106 the 109 proposition 120 that 124 all 127 men 130 are 133 created 140 equal 145. Now 148 we 150 are 153 engaged 160 in 162 a 163 great 168 civil 173 war 176 testing 183 whether 190 that 194 nation 200 or 202 any 205 nation 211 so 213 conceived 222 and 225 so 227 dedicated 236 can 239 long 243 endure 248. We 251 are 254 met 257 on 259 a 260 great 265 battlefield 276 of 278 that 282 war 285.

FIG. 5. SECTION OF THE UPPER END OF THE GETTYSBURG HANDWRITING SCALE

of the Gettysburg Address until they are familiar with them. They are then provided with pen, ink, and ruled paper. At a given signal they begin to write and continue for two minutes. The papers are then collected for scoring.

Scoring the Papers—The samples are scored on quality by passing each paper in turn along a copy of the scale until a point is found where the quality of the sample and the quality of the scale agree. The accompanying

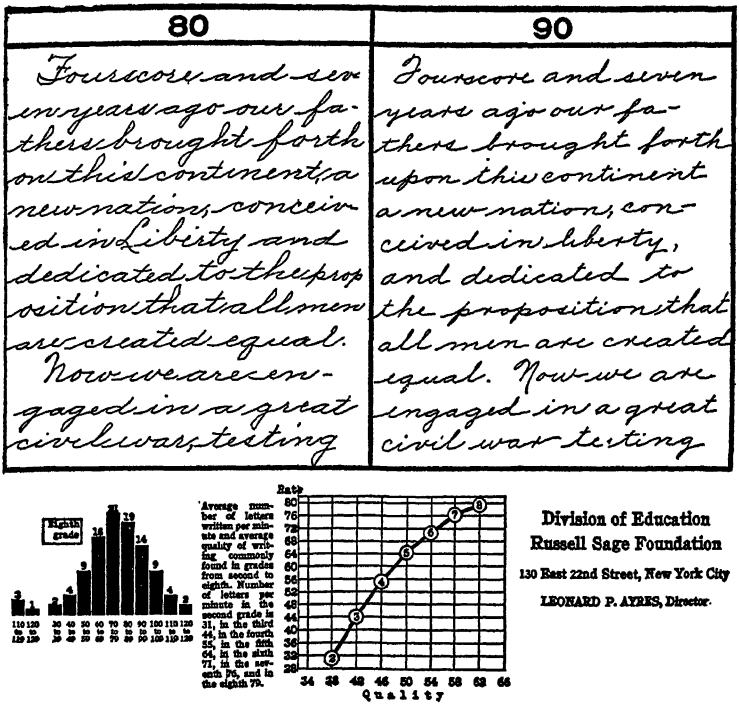


FIG. 6. LOWER ENDS OF THE GETTYSBURG HANDWRITING SCALE

score on the scale is given as the grade for the sample of handwriting. In case the sample seems to fall between two points on the scale intermediate scores may be given. In grading quality, differences in style are to be disregarded. The rate of writing is determined by finding the average number of letters written per minute during the two minutes writing period.

Norms—Norms for both speed and quality are given as follows:

AYRES HANDWRITING NORMS

Grade	2	3	4	5	6	7	8
No. of letters written per min.	31	44	55	64	71	76	79
Quality score	38	42	46	50	54	58	62

From this table it may be seen, for example, that a fourth grade child should write at the rate of 55 letters per minute with a quality of 46 on the Ayres Scale.

Function of the Scale—The use of this scale makes the grading of handwriting as objective as possible. As has already been indicated a relatively small amount of practice makes the average classroom teacher very proficient in the use of the scale. It is simple to administer and the norms are very reliable. Some teachers may object to making legibility the basis for scoring handwriting. If so, the Thorndike Scale should be used. But practically, as least, no other criteria are so important or satisfactory as legibility as a basis for merit.

The Thorndike Scale for the Handwriting of Children in Grades Five to Eight

Description of the Scale—This scale was constructed by Professor E. L. Thorndike and consists of 29 samples of handwriting ranging in quality from 4 to 18. There is a sample for each step on the scale and in many cases more than one sample for a step, representing different styles of handwriting. Steps in the scale represent differences in general merit of handwriting as described later.

Derivation of the Scale—The material for the qualities from 5 to 17 of the scale was taken from actual samples of the handwriting of children. The sample for quality 4 was artificially constructed and for quality 18 was

taken from a copy book. There were about 1,000 samples in all. These samples were rated by from 23 to 55 judges on the basis of general merit [6] of handwriting. Slant, style or other special factors were not taken into consideration in the rating.

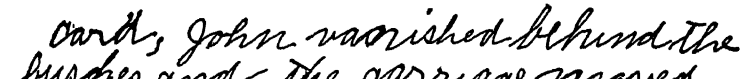
While there is no sample that was ranked as zero, this point was defined "roughly as handwriting recognizable as such but of absolutely no merit as handwriting." This zero point is of theoretical interest, at least, for as Thorndike has pointed out three things are necessary for any kind of measurement: (1) a zero or beginning point, (2) an ending point and (3) a unit of measure. Thorndike uses a rather complicated mathematical method of arriving at his unit of measure but so constructs it that the difference between sample 4 and sample 5, for example, is the same as the difference between sample 17 and sample 18; so that sample 8 is just twice as good as sample 4. "The unit of the scale equals one-tenth of the difference between the best and the worst of the formal writing of 1,000 children in grades 5 to 8." [7]

Method of Using the Scale—The scale is to be used by comparing the specimens of handwriting to be measured with the samples in the scale and assigning a score on the basis of this comparison from the scores given on the scale. Fractional scoring between the units of the scale is allowed. If it is desired to grade on the basis of 100, the Thorndike score may be transposed to this basis by multiplying it by 5.5. For example, a score of 12 on the

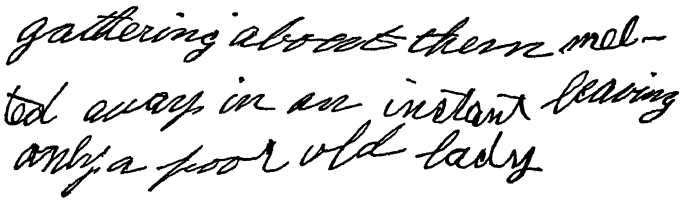
[6] The authors are aware that Monroe and Hines say the rating was on the basis of beauty, legibility, and general merit, but they can find no authority for such a statement.

[7] See "Handwriting," E. L. Thorndike, *Teachers College Record*, Vol. II, No. 2 (March, 1910).

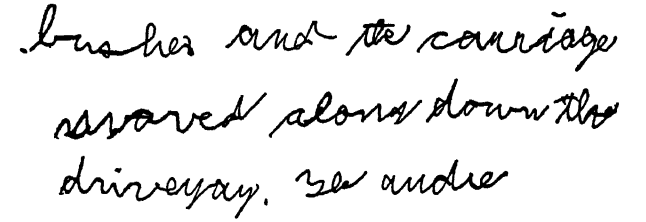
Quality 7. Sample 126



Quality 6. Sample 12



Quality 5. Sample 6



Quality 4. Sample 121

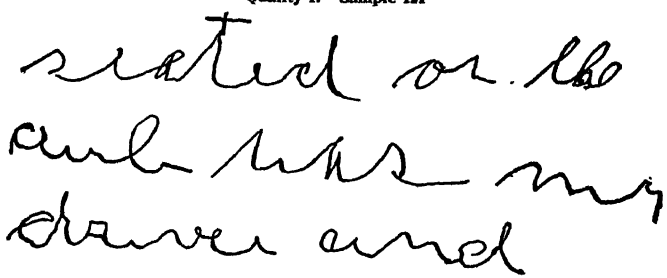


FIG. 7. A SECTION OF THE THORNDIKE HANDWRITING SCALE

Thorndike scale is equivalent to a score of 66 on the scale of 100.

We have already considered the problem of the training of the teacher in the use of a handwriting scale. Thorndike has provided a valuable means whereby the teacher may improve her scoring in handwriting by supplying fifty samples of handwriting with the true values of the specimens graded by the Thorndike scale.[8] The teacher may practice grading these specimens and by comparing her scores with the true scores may determine the direction and amount of her errors in scoring. Thorndike says that an average competent teacher who is without training in the use of the scale will make an error of .9 (4.95 on the scale of 100) of a step in judging a sample. Practice on the fifty specimens with knowledge of the results should lower this error.

Function of the Scale—The Thorndike Scale is well suited for the purpose for which it is intended, that is, as an aid to the teacher in scoring handwriting on the basis of general merit. The scale is not diagnostic as it does not indicate the elements of merit or demerit. Norms have not been derived for the different school grades.[9] Speed of writing is not taken into account in this scale. Its purpose is to present an objective standard as the basis for grading handwriting.

[8] "Teachers' Estimates of the Quality of Specimens of Handwriting," by E. L. Thorndike, *Teachers College Record*, Vol. XV, No. 5 (Nov. 1914).

[9] Starch gives norms for the Thorndike Scale based on a study of 6,000 pupils in 28 schools. Daniel Starch, *Educational Measurements*, p. 83.

Grade	1	2	3	4	5	6	7	8
Speed	20	31	38	47	57	65	75	83
Quality ...	6.5	7.5	8.2	8.7	9.3	9.8	10.4	10.9

The Gray Standard Score Card for Measuring Handwriting

This is not a test or scale but, as its name indicates, a card for recording certain qualities of handwriting. It is constructed on the same general principle as score cards used in judging stock and fruit and is to be used in the same way and for the same general purpose. It is to point out the good and the poor qualities in handwriting. There are nine separate qualities listed on the card. The highest possible score varies from 3 for heaviness of line to 26 for general form; 100 points represents a perfect score.

Like the Freeman Scale this score card is valuable for pointing out the different qualities that go to make up handwriting and the relative importance of each. Too often the pupil and even the teacher think only in terms of general merit without any very definite idea of what constitutes merit. The teacher may make use of the score card for grading her pupils in handwriting and in this way it really becomes a diagnostic scale.

The Courtis Standard Practice Tests in Handwriting

Description of the Tests—These are not tests in the ordinary sense of the word but are a combination of tests and practice exercises in handwriting. The tests and the method of using them are described in a Student's Daily Lesson Book and a Teacher's Manual. The Daily Lesson Book provides each child with copies for practice lessons and lesson helps to go with each copy. The Teacher's Manual describes the tests, how the pupils are to use the tests, how the teacher is to help the children in the

use of the tests, how to measure progress through the use of the tests, and what use the teacher should make of the results.

Derivation of the Tests—These tests were devised by S. A. Courtis and Lena A. Shaw of Detroit as a result of three years' experimentation in handwriting. A series of twenty lessons chosen from a text book in business writing was arranged so that the easiest forms came first and the more complex forms followed. The lessons progressed from simple words to more difficult words and phrases, then whole sentences and paragraphs. These lessons were supplemented later by others and the material rearranged in a more logical form.

In a later arrangement, the material for the first lessons was determined by a study of the relative frequency with which various letters of the alphabet occur in everyday usage, except that this principle was somewhat modified because of differences in difficulty of the formation of some letters. Standard rates and qualities for the Courtis Tests were derived from a study of 1,000 samples of handwriting. The quality of these specimens was measured by, and the norms given in terms of, the Ayres Scale.

Method of Using the Tests—"On the first day a research test is given in order to find out what children need drill, what kind of drill and how much, and whether there are children in the class who would not profit by working on the practice tests. Those who fail to pass the research test begin on Lesson One, and each child must continue on that lesson until he has written it fast enough and of a quality up to standard for his grade before he can pass to the next lesson. The result is that

children work on those lessons on which they need to work. The child who can progress at the correct rate does not need the help of his teacher. The child who goes too slowly receives the individual attention of the instructor. Provision is made for the teacher to discover his slow progress, or lack of progress, and the weakness in his writing. She then helps to remedy them.

"The child enters his own record and graph after every day's work. He scores his own paper. As a result of these two operations on his part he learns to judge for himself why his work is not so good as it should be and how much more rapidly he should do his work in order to make it equal, or surpass, that which is set as the standard for his grade.

"The children who do not need drill from the beginning of the work or those who finish the series of lessons in a very short time may be excused from practice work in handwriting for that grade, or they may be given the standard for the next grade, whichever method seems advisable in the particular school system in which the child is working." [10]

In general each day's work is to be divided into three parts: (1) special practice, 5 minutes, (2) testing, 5 minutes, (3) scoring and recording, 5 minutes. The purpose of the practice tests is "to teach the children to teach themselves to write well." As already stated the pupils are to grade their own specimens of handwriting. Quality is scored by the use of the Ayres Scale. Rate is the number of letters written in three minutes. The following norms are given:

[10] Bulletin No. 1, Courtis Standard Practice Tests in Handwriting, p. 9.

COURTIS HANDWRITING NORMS

Grade	3		4		5		6		7		8	
	low	high	low	high	low	high	low	high	low	high	low	high
Standard Rate	40	46	52	58	62	66	69	72	73	78	80	82
Standard Quality ...	45		50		55		60		65		70	

Thus a pupil in the high sixth grade should write 72 or more letters in three minutes and the quality should be at least 60 on the Ayres Scale.

Value of the Tests—In order to test the value of the use of the tests, Courtis arranged two groups of pupils, one of which used the practice tests and the other received the ordinary instruction in handwriting. The two groups were so arranged that their handwriting was about equal at the beginning of the experiment. The differences of the two groups at the end of the school year is shown in the accompanying graph.

"In every class, except one, the pupils who used the Courtis Standard Practice Tests in Handwriting made a greater percentage of gain than those who studied penmanship without them, the least gain being a little less than 20% better, and the greatest gain being as much as 380% better." The beginning teacher, or the grade teacher without a supervisor, can make especially good use of such a carefully constructed teaching device as these practice tests in handwriting provide.

Materials Needed

1. Ayres Handwriting Scale, Gettysburg Edition for grades 2 to 8, Russell Sage Foundation, New York City, price 10 cents, or Public School Publishing Co., Bloomington, Ill., price 18 cents.

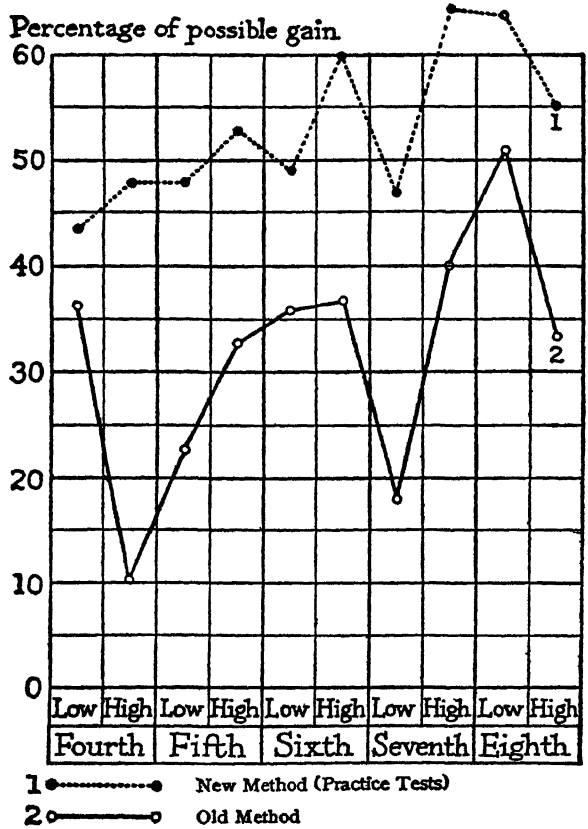


FIG. 8. FROM COURTIS BULLETIN, NO. 1, P. 9

2. Courtis Standard Practice Tests in Handwriting. Specimen Set including Student's Daily Lesson Book, Student's Daily Record Card, Teacher's Manual, Ayres Handwriting Scale, Gettysburg Edition (2 copies), all for grades 3 to 8, World Book Co., Yonkers on Hudson, price 45 cents.
3. Freeman Chart for Diagnosing Faults in Handwriting for grades 2 to 8. Houghton Mifflin Co., New York City, price 30 cents.

4. Gray's Standard Score Card for Measuring Handwriting for grades 2 to 8. Public School Publishing Co., Bloomington, Ill. Price for sample 15 cents.
5. Thorndike Handwriting Scale for grades 2 to 8. Bureau of Publications, Teachers College, Columbia University, New York City, price 15 cents.

Selected References

Ayres, L. P., *A Scale for Measuring the Quality of Handwriting of School Children,* Russell Sage Foundation, New York City, Bulletin No. 113.

Freeman, F. N., *The Teaching of Handwriting,* Houghton Mifflin Co., 1914.

Freeman, F. N., "The Handwriting Movement," *Supplementary Educational Monographs,* Vol. II, No. 3.

Koos, L. V., "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools," *Elementary School Journal,* February, 1918.

Thorndike, E. L., "Handwriting," *Teachers College Record,* March, 1910.

Thorndike, E. L., "The Measurement of Ability in Handwriting," *Teachers College Record,* November, 1914.

CHAPTER VII

READING

Reading is the most fundamental subject of the elementary school curriculum. With this in mind, let us first consider some of the psychological problems involved in reading. Very early in the life of a child it learns to connect certain sounds with certain objects as a result of ordinary experiences. As a result of further experiences and observations the child learns to connect words into spoken discourse. It learns to express its wants, feelings and ideas by the use of words and sentences. This process is well developed in the child before there is any attempt to teach it to read. Reading for the beginner is the process of connecting written or printed symbols with certain sounds. As taught by modern methods, these sounds are the words, the meaning of which the child already knows. The problem for the child and the teacher is, therefore, a double one, that of connecting symbols and words with their spoken sounds; and also that of getting meaning from the printed or written words. The first is a mechanical process and requires long patient drill. The second, although dependent upon the first, is ultimately by far the more important since the purpose of reading is the gaining of meaning from the printed page.

In the beginning the words must be spoken by the

child. This is the period of oral reading. Later the words may and for the most part should be read silently by the child. Here the mechanics drop into the background and reading becomes a process of thought-getting. This thought-getting may consist in the mere understanding of the meaning of the printed words, it may consist in the understanding of sentences or in the interpretation of paragraphs or whole selections, or it may be a combination of all these factors.

In emphasizing the importance of reading, Dr. W. A. Schmidt has pointed out the fact that more than one-fourth of the time in our elementary schools is devoted to the formal study of reading. But much more significant than this is the fact that it is by means of reading that the pupil, as well as the adult, obtains the larger share of his information. It is very evident that the pupil's success in history, geography, literature and hygiene depends almost exclusively on his ability to extract meaning from the printed page. Too often the teacher does little more than quiz the pupil on his reading, but at the best she only explains, elaborates and interprets the lesson assignment. Even in a subject like arithmetic, reading is a very important factor. Dr. Paul Terry has shown by photographs of the eye movements of pupils while reading problems in arithmetic, that much of the difficulty in this subject lies in poor reading. Manifestly a pupil can not solve a problem if he does not understand what the problem means.

What has been said of reading in the grades applies much more truly to the high school and college student. Many of the failures in high school can be traced directly to poor reading rather than to a lack of intelligence. With

the great increase in the amount of reading demanded of the high school pupil comes the added demand for rapid, efficient, silent reading. If such reading habits have been developed the pupil is prepared for his tasks. If he be poorly prepared in the mechanics of reading his chance of success is limited. It is the verdict of many high school teachers that the average pupil is not well prepared. In fact very few are as well prepared as they should be. This condition may be due either (1) to an inability to comprehend the printed page, (2) the formation of habits of slow reading or (3) both of these combined.

By the methods too commonly used in teaching children to read, thought-getting has not been emphasized as the aim in reading. The child has been taught that word pronouncing is the end. This has sometimes come from improper training in phonics and prolonged emphasis on oral reading. The result is that the pupil does not develop the habit of reading for meaning.

Speed in reading is seldom sufficiently emphasized in reading. How often have we heard a teacher say, "Now, don't read too fast, John." This caution may be justified in oral reading when the pupil tries to read more rapidly than he can articulate properly. But there is little danger in too rapid silent reading. In fact most of us could read a half faster without loss in comprehension. There has been an investigation [1] of the effect of reading at normal, rapid and slow speeds on ability to recall what was read. It was found that there was very little difference in the amount recalled by any of the

[1] A. R. Gilliland, "The Effect of Rate of Silent Reading on Ability to Recall," *Journal of Educational Psychology*, November, 1920, Vol. XI, p. 474 ff.

three methods of reading. When the gain in time is taken into consideration, reading at the most rapid rate was much more efficient. The elementary school pupils gained about one-fourth by the rapid reading and the high school and college students even more.

In this connection it is worth noting that most of us by a little judicious practice could materially increase our rate of reading. Dr. E. B. Huey,[2] after enumerating a number of experiments in which speed of reading was greatly increased without loss in comprehension, says that he doubled his own reading rate as a result of practice in rapid silent reading.

When comprehension is poor and the rate of reading slow, a pupil is seriously handicapped and it is only by prolonged and concentrated effort that even a mediocre amount can be accomplished. If this effort is not forthcoming, the pupil is doomed to failure.

Tests in reading serve two general purposes. The first purpose is to measure success or failure during the period while the pupil is learning to read. By means of diagnostic tests defects in the learning process may be detected and corrective measures applied. The second use for reading tests comes later in the pupil's career. As has already been pointed out, one of the common causes for poor work in the upper grades and in high school is poor reading. Poor students should be given a reading test to determine whether this is the cause of failure. If so, methods such as those suggested by Judd[3] should be used to remedy these defects. In this connection it

[2] E. B. Huey, *The Psychology and Pedagogy of Reading*, page 180.
[3] C. H. Judd and others, "Reading, Its Nature and Development," *Supplementary Educational Monographs*, University of Chicago, Chapters V and VI.

should be noted that general comprehension and speed tests, rather than the more diagnostic tests, should be used for this purpose.

Many tests have been constructed for measuring ability in reading. One of the best known of these tests is designed to measure ability in the mechanics of oral reading. Other tests measure ability to understand the meaning of words, sentences, or whole selections. Many different methods of recording responses are used. In some of the tests the pupil is to draw a line under or around certain words, in others he is called upon to reproduce the story orally or in writing. In still others the pupil is to answer certain questions based on what has been read.

The teacher may become confused by the large number of tests and may ask which of these tests to use. This depends largely upon what phase of the reading process is to be tested. If the tester desires to find out whether the pupils are prepared in the mechanics of oral reading, she will use the Gray Standardized Oral Reading Tests. If she is concerned with the ability of the pupils to understand the meaning of words, the Pressey or Thorndike Visual Vocabulary Scale will be used, the Pressey Test for the lower grades and the Thorndike in grades three to second year high school. If the teacher desires to exclude the influence of handwriting on reading, the Kansas Silent Reading Test, the Burgess Scale for Measuring Ability in Silent Reading, the Courtis Silent Reading Tests, or the Monroe Standardized Reading Test should be used. Some of these are much more comprehensive than others. Some emphasize thought-getting

while others involve interpretation and reasoning. The Gray Silent Reading Tests require the pupils to reproduce the story and answer a list of questions based on the story. The Haggerty tests measure word meaning, sentence meaning and paragraph meaning. The teacher, therefore, must decide what phase of the reading process she wishes to test and select the test accordingly.

The Gray Standardized Reading Paragraphs

Description of the Test—This oral reading test was devised by Dean W. S. Gray of the University of Chicago. It consists of twelve short paragraphs ranging in difficulty from very easy material in the first paragraph for pupils of the lower grades to paragraphs difficult enough to tax the ability of high school pupils. This increased difficulty consists largely in the use of longer and more unusual words. The purpose of the test is to measure ability in the mechanics of oral reading. It is an individual test.

As the pupil reads one after another of the paragraphs the tester measures the time required to read each paragraph and records the errors in reading. Six types of errors are recorded. These errors are:

1. Gross Errors—Total mispronunciation of words.
2. Minor Errors—Partial mispronunciation, as wrong vowel sounds or accent.
3. Omissions—Leaving out a word in the reading.
4. Substitutions—Reading another word instead of the one in the text.
5. Insertions—Adding words not in the text.
6. Repetition—The rereading of two or more words after they have already been read.

In order to facilitate the recording of errors a method of recording is suggested as illustrated.

The sun pierced into my large windows. It was the opening of October, and the sky was of a dazzling blue. I looked out of my window and down the street. The white houses of the long straight street were almost painful to the eyes. The clear atmosphere allowed full play to the sun's brightness.

"If a word is wholly mispronounced, underline it, as in the case of 'atmosphere.' If a portion of a word is mispronounced, mark appropriately, as indicated above: 'pierced' pronounced in two syllables, sound long a in 'dazzling,' omitting the s in 'houses' or the al from 'almost' or the r in 'straight.' Omitted words are marked as in the case of 'of' and 'and'; substitution as in the case of 'many' for 'my'; insertions as in the case of 'clear'; and repetitions as in the case of 'to the sun's.' Two or more words should be repeated to count as a repetition."

"Each pupil should be allowed to continue reading until he makes at least the following number of errors in each of two paragraphs: 5 errors or more in 40 or more seconds, or 7 or more errors in case the paragraph is read in less than 40 seconds."

Scoring the Results—A pupil's score is a combination of his rate of reading and the number of errors made. A table is provided on the score sheet for determining the combined score. For example, if the pupil requires 26 seconds to read a paragraph and makes 3 errors, his score

is 2. A somewhat complicated method of obtaining the final individual or class score is fully described on the score sheet. Norms have been carefully worked out for all of the school grades.

Sample Paragraphs from the Gray Standardized Reading Tests

I

A boy had a dog.
The dog ran into the woods.
The boy ran after the dog.
He wanted the dog to go home.
But the dog would not go home.
The little boy said,
"I cannot go home without my dog."
Then the boy began to cry.

6

The part of farming enjoyed most by a boy is the making of maple sugar. It is better than blackberrying and almost as good as fishing. One reason why a boy likes this work is that someone else does most of it. It is a sort of work in which he can appear to be very industrious and yet do but little.

12

The hypotheses concerning physical phenomena formulated by the early philosophers proved to be inconsistent and in general not universally applicable. Before relatively accurate principles could be established, physicists, mathematicians, and statisticians had to combine forces and work arduously.

NORMS FOR GRAY STANDARDIZED READING PARAGRAPHS

School grade....	1	2	3	4	5	6	7	8
Grade score.....	31	42	46	47	48	49	47	48

A fifth grade child therefore should make a score of 48 on the test. The apparent lack of improvement in the score for the different school grades is due to the different credits for the first paragraph.

Function of the Test—This is one of the most objective of the educational tests. While it does not measure all the factors in oral reading, such for example as expression or thought-getting, the ones that are measured are definite. The test is highly diagnostic in the sense that the records show not only the number but also the types of errors made in oral reading. The teacher by reference to the score sheet can determine the factors in which her pupils are below average in their reading. If this is a matter of only a few pupils these pupils should receive special drill, or if it is general, but of only one or two types of errors, the teacher should consider whether or not she has been giving enough attention to these points. If the class is uniformly low other causes should be sought. The pupils may have had poor previous preparation, they may not have the general mental ability to profit by instruction, possibly not enough time is being devoted to reading, or, the method of teaching the subject is at fault and should be changed. These problems should each receive careful consideration and study. Enough has been said to indicate the value of such a test as this for the teacher and how she may make use of the results of the test in improving her work in reading.

The Pressey First Grade Reading Vocabulary Test and First Grade Reading Scale

Description of the Tests—The Vocabulary test consists of twenty-seven sets of syllables with five syllables in each set. Each set contains one real word and the other four are nonsense syllables. The pupils are given six minutes in which they are to draw a circle around the real word in each set of syllables.

Part of the Word Test of the Department of Psychology, Ohio State University. Devised by L. W. Pressey and Viola Grant.[4]

THE PRESSEY FIRST GRADE READING SCALE

1. is the you a said
2. do we come are ball
3. baby one with that have
4. good was on this his
5. mouse has your match bird
6. oh give for mother ran
7. fly very water as milk
8. home help blow some girl
9. three won't be name will
10. rabbit bumblebees over shall bear

The First Grade Reading Scale is composed of a word test and a sentence test. The word test consists of twenty-five columns of five words each. The pupils are directed to draw a line around a certain word in each column. Each succeeding list of words is somewhat more difficult than the preceding. The sentence test is similar except that the fifteen columns are composed of

[4] Published by the Public School Publishing Co., Bloomington, Ill.

short sentences instead of disjointed words. There is no time limit for these tests. Alternate forms of both the Vocabulary and Reading Scale are available.

Function of the Tests—The general purposes of these tests are very similar. They constitute measures of recognition of words in the reading vocabulary of first grade children. These tests fill the need of a measure of the knowledge of words out of context for the first grade [5] very much the same as the Thorndike Visual Vocabulary Scale does for the upper grades.

The Thorndike Visual Vocabulary Scale

Description of the Scale—The scale consists of a graded series of words which the pupil is to classify according to certain specified groups. This classification is accomplished by placing a letter or word under each word. For example, the letter "F" is to be written under every word that means flower, the letter "A" under every word that means animal and the word "Bad" under every word that means something bad to be or do.

The words are arranged in groups and these groups are assigned numerical values on the basis of their difficulty. The first group in scale A contains five words with a value of 4. The last group contains three words with a value of 11. A preliminary test, sometimes called a shock absorber, precedes the test proper to familiarize the pupil with the nature of the test. There are four forms of the scale: A2z, A2y, Bx, and By. The four scales are of approximately equal difficulty and may be used from grade three to second year high school.

[5] A reading test containing a measure of vocabulary also constitutes a part of the Pressey Second Grade Attainment Scale.

Method of Using and Scoring—After the preliminary test has been given the child is furnished a copy of the scale. The directions are printed on the scale. No time limit is set; hence, the child is allowed to work until he finishes. The pupil's score is the number of the highest numbered group of words in which he makes not more than a single error. The numbers roughly indicate the average effect of this number of years of training on the vocabulary of the child.

Function of the Scale—This scale is especially valuable as a measure of the pupil's knowledge of the meaning of words out of context. While word knowledge is really not a part of the reading process proper, it is such an essential basis for success in reading that the Thorndike Scale is here listed as a reading test. If a pupil is doing poorly in his reading the teacher may well use such a test as this to determine whether or not the pupil's difficulty is with vocabulary. If so, the pupil should be drilled on word meanings. If the pupil makes a good score in the Visual Vocabulary Scale and still is a poor reader, some of the tests that measure the mechanics of reading or sentence and paragraph meanings should be used in order to locate his difficulty.

First Half of the Thorndike Reading Scale D. Word Knowledge or Visual Vocabulary

Write the letter W under every word that means something about *war* or *fighting*.
Write the letter B under every word that means something about *business* or *money*.
Write the letters CHU under every word that means something about *church* or *religion*.
Write the letter R under every word like *father* or *wife* that means something about *relatives* or the *family*.
Write the letters COL under every word that means a *color*.
Write the letter T under every word like *now* or *then* that means something to do with *time*.
Write the letter D under every word like *here* or *north* that means something about *distance* or *direction* or *location*.
Write the letter N under every word like *ten* or *much* that means something about *number* or *quantity*.

4x. camp, flag, west, mother, two, general, green, troops, south, fort
4½x. gray, cousin, pink, uncle, yellow, hour, pay, aunt, early, commander
5x. marriage, defeat, many, afternoon, guard, buy, captive, military, relation, late
6x. hymn, defend, across, merchant, noon, forty, conquer, dagger, profit, tuesday
6½x. month, dozen, fortress, cavalry, tax, bishop, below, october, million, owe
7x. fortification, ownership, there, year, june, half, scarlet, soon, november, beneath

The Monroe Standardized Silent Reading Test (Revised)

Description of the Test—This test was devised by Professor W. S. Monroe, director of the University of Illinois Bureau of Educational Research. It is a test of

Two Paragraphs from the Monroe Standardized Silent Reading Test. Revised from Test II for Grades 6, 7, and 8, Form 3

537 11. Beside our house was a little hut where a holy man lived
550 in charge of an adjoining shrine, earning money for himself
562 and for the shrine by polishing little pieces of marble as mementos for visitors.

573 Draw a line under the word which best describes this holy man.

585 *industrious good foolish lazy sad*

590 12. Nanook, once so full of life, now knew perfectly well
602 that it was all over with him. Head and tail down, the picture
615 of resigned dejection, he stood like a petrified dog.

622 Draw a line under the word which best describes the dog Nanook.

634 *angry frightened active hungry down-hearted*

both speed and comprehension. The reading material consists of a series of short paragraphs, each paragraph followed by a list of words. The pupil is directed to underline one of these words, the information as to which word is to be underlined is contained within the paragraph. Four minutes is the time allowed in which the pupil is to read and underline as many words as he can. Test I is for grades 3, 4, and 5 and Test II is for grades

6, 7, and 8. Each test is issued in three different forms of equal difficulty.

Scoring the Test—The comprehension score is the number of exercises the pupil answers correctly in the four minutes allowed for the test. This score is then transferred into an Accomplishment Age score by means of a table given in the Teachers' Handbook. The rate of reading may also be transferred into an Accomplishment Age score by reference to the same table. The author suggests that these two accomplishment scores may be averaged as a single measure of silent reading ability.

Norms are given for comprehension score and rate score in terms of achievement age. Three norms—one for rural schools, one for city schools and the other a general norm—are given for each grade. These norms were based on 55,000 scores.

NORMS FOR MONROE STANDARDIZED SILENT READING TEST (REVISED)

School Grade.....	3	4	5	6	7	8
COMPREHENSION	yrs.mo.	yrs.mo.	yrs.mo.	yrs.mo.	yrs.mo.	yrs.mo.
Rural	7-11	9- 0	10- 0	11- 1	13- 1	13-10
Cities	8- 2	9- 9	11- 6	12- 6	13- 5	14- 6
General	7-11	9- 4	10- 9	12- 0	13- 3	14- 2
RATE						
Rural	7-11	9- 1	10- 2	11- 7	13- 0	14- 6
Cities	8- 2	9- 7	10-10	12- 2	13- 4	15- 4
General	7-11	9- 5	10- 7	11-11	13- 2	14-10

Function of the Test—This test is easy to give and requires only a small amount of time. It requires little writing on the part of the pupil and the grading is not difficult. As has been pointed out in another place, a test should, so far as possible, isolate one factor, or at least only a few concrete factors and measure them. In the

case of reading, both rate and comprehension can be measured at the same time. But in this test the answer depends upon reasoning ability as well as reading ability. In so far as this is true it becomes a test of [redacted] rather than one of reading ability. The method of measuring rate of reading may also be questioned. Rate is measured by the amount of work completed in four minutes. Part of this time is devoted to reading and another part to underlining the answers. A pupil might read very rapidly and take a long time in the underlining and receive a low score in the rate. Monroe's method of averaging comprehension score and rate score as a measure of reading ability is unique.

The Kansas Silent Reading Test

Description of the Test—This test was devised by Professor F. J. Kelly while director of the training school of the Kansas State Normal School. It is composed of a series of short paragraphs, each paragraph making a statement followed by a question or giving directions to be followed and a line upon which the answer is to be written or a series of words, one of which is to be underlined or encircled. The paragraphs are arranged in order of difficulty and a value ranging from 1 to 52 is assigned to each paragraph. Five minutes is allowed for the test and the pupil's score is the sum of the values of the paragraphs which were completed correctly.

There are three separate tests. Test I is for use with grades 3, 4, and 5, test II is for grades 6, 7, and 8 and test III for grades 9, 10, 11, and 12. Median scores have been determined from the use of the tests in nineteen cities. They are as follows:

State Normal School,
EMPORIA, KAN.
Bureau of Educational Measurements and Standards.

Put Pupil's Score Here.

THE KANSAS SILENT READING TEST.

Devised by F. J. Kelly

FOR

Grades 6, 7 and 8.

Value 2.6.

No. 8.

Here are two squares. Draw a line from the upper left-hand corner of the small square to the lower right-hand corner of the large square.

Value 3.0.

No. 9.

A farmer puts one-half the hay from his field into the first stack, then two-thirds of what is left into a second stack, and the remainder into a third stack. Which stack is the largest?

Value 3.9.

No. 10.

Below are two squares and a circle. If the circle is the largest of the three, put a cross in it. If one square is smaller than the circle, put a cross in the large square. If both squares are smaller than the circle, put a cross in the small square.

FIG. 9. TEST II. FROM THE KANSAS SILENT READING TEST

THE KANSAS SILENT READING TEST

School Grade..	3	4	5	6	7	8	9	10	11	12
Median Score..	6.0	9.9	13.7	13.4	16.5	18.8	22.9	25.8	26.0	.8

Function of the Test.—This test requires little handwriting, is easy to give and to score, and takes only a short time to give. For these reasons it is a very practical classroom test. Like the Monroe Test many of the answers depend upon more than ability to comprehend the paragraph. Some require mathematical ability and others call for geographical and historical information not specifically contained within the paragraph. We must define reading ability very broadly if we demand that the pupil furnish information in his answers other than that contained in the material read. As might be expected the Kansas Silent Reading Test has been found to correlate very highly with scores of general intelligence.

The Thorndike-McCall Reading Scale

Description of the Scale—This silent reading scale is made up of a series of paragraphs each followed by a list of questions. There is a preliminary paragraph for practice and nine other paragraphs ranging in difficulty from the first which is easy enough for a second grade pupil to one difficult enough for a high school senior. There are 35 questions based upon the paragraphs. Thirty minutes is allowed for the test and during this time the pupil may read and reread the paragraphs as many times as he likes. The papers are graded by determining the number of questions answered correctly. This number is transformed

Sample Paragraph from the

Thorndike-McCall Reading Scale for the Understanding of Sentences—Form 4

Write your name here.....
School..... *Grade*..... *Date*.....
How old are you?..... *When is your birthday?*.....

This is to be a reading contest. You will read paragraphs like this one, and answer questions like those you see below. Answer every question you can. If you come to a question you can't answer skip it and go on. Go back to it later. If you finish before you are told to stop, go back and make sure you have made no mistakes. When possible the answers to the questions must be found in the paragraph. You may read the paragraph as many times as you need to. You will have enough time but don't waste it. Play fair. Don't look at anyone else's paper. You will be told your score later.

Read this and then write the answers. Read it again if you need to

In August, Arthur and his Cousin Kate went in the train to visit their grandfather, Mr. Peters, at Oak Farm. They played in the brook, picked blackberries, and hunted for eggs in the barn. They played with Bob Peters and Nan Allen. Bob was nine years old; Nan was eleven.

5. How old was Nan?.....
6. Which was older, Bob or Nan?.....
7. Does the story tell how old Kate was?.....
8. Does the story say that Bob and Nan went in the train?.....

.....

Do the next page.

into what is called a T score [6] by means of a table which is furnished with the scale. Ten equivalent and inter-

[6] See Wm. A. McCall, *How to Measure in Education*, Macmillan, 1922, Chap. X, pp. 272-306, for a discussion of the T scale.

changeable forms of the scale have been constructed. The following norms are given:

THE THORNDIKE-McCALL READING SCALE

School Grade.....	2a	2b	3a	3b	4a	4b	5a	5b
T Score Norm.....	26	30	33.7	37.3	39.6	41.8	44.9	48.0
School Grade.....	6a	6b	7a	7b	8a	8b	9a	9b
T Score Norm.....	50.9	53.7	56.0	58.3	59.6	60.9	61.5	62.1
School Grade.....	10a	10b	11a	11b	12a	12b		
T Score Norm.....	62.9	63.6	64.5	65.4	66.8	68.1		

Function of the Scale—The subject matter of this test is the ordinary sort of material that a child might be expected to read. The answers to the questions are based directly on the material contained in the paragraphs. Not much writing is required of the pupils and yet the questions are not of the "yes" and "no" type. The test does not measure rate of reading except indirectly. The alternative forms of the test make it possible to make frequent retests of pupils in order to measure progress. In general, this is a very satisfactory test of comprehension in silent reading.

The Burgess Scale for Measuring Ability in Silent Reading

Description of the Scale—The Burgess Scale consists of a series of twenty pictures each followed by directions. The pupils are told that each paragraph "tells them to do something to the picture above it with their pencils." They must read carefully to make sure what they are to do. The paragraphs are to be read and marked in order, starting at the top and working down. The pupil must

do as many as he can in the five minutes allowed for the test.

Scoring the Papers—Every paragraph is counted as correct in which the marking of the picture, no matter

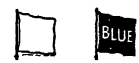
1. This naughty dog likes to steal bones. When he steals one he hides it where no other dog can find it. He has just stolen two bones, and you must take your pencil and make two short, straight lines, to show where they are lying on the ground near the dog. Draw them as quickly as you can, and then go on.

5. Have you ever seen such a strange bird? He is hard to find because he sleeps in the woods during the day and does not come out until night. Take a pencil and tell people what the bird's name is by writing the word OWL, with a capital O, under the books on which the bird is standing.

2. This man is an Eskimo who lives in the far north where it is cold. There has just been a big storm, and all the ground is white with snow. The man has been walking and has made many footprints in it. With your pencil quickly make four of these in the snow just behind him.

6. This small chap is afraid to start for school. The teacher will scold unless he brings his books; but the big owl is sitting on them. Grasp your pencil bravely and cross the owl out of the previous picture with two black lines, so that the child can rescue his belongings. Remember not to use more than just two lines.

3. This book is lying on the desk, but it is hard to make it stay open. With your pencil draw a single straight line to represent a ruler lying across the book to hold the pages open. Be sure to make the line from one side to the other, across the book, instead of making it go up and down.



7. These two flags are used as signals to give notice of changes in the weather. The white flag means fair; so you may now take your pencil and make a capital F under the white flag, to stand for fair. The blue flag means storm; so make a capital S under the blue one.

FIG. 10. SAMPLE PARAGRAPHS FROM THE BURGESS SILENT READING SCALE

how crude it may be, exactly follows instructions. The pupil's score is the number of paragraphs marked correctly. This score may be transferred into a score on the basis of 0 to 100 for any school grade by reference to the accompanying table.

Test Score	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Grade 3.....	0	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..	..	..	..	..	..
" 4.....	0	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..	..	..	..
" 5.....	0	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..	..	..
" 6.....	0	2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..	..
" 7.....	..	0	2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..
" 8.....	..	..	0	2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100

The scores on this table are so constructed that the average pupil of any grade should receive a score of 50. Above average is denoted by a score of more than 50 and below average by a score below that mark. These scores are for February 1st. For other months of the school year certain additions or subtractions are to be made from the scores.

Function of the Scale—The author of this scale states that it was so devised that it should be free from four fundamental limitations of the ordinary scale. (1) It is expected that this scale will measure ability in reading and not other abilities such as handwriting or English composition. (2) The scale is uniform and relatively equal in difficulty throughout. It measures one and only one ability. (3) It is easy to administer and score. (4) The score for each school grade is available for comparison with the standing of other children.

In so far as this test conforms to these four principles it is superior to most other scales. But some of the paragraphs call for a drawing ability not possessed by some children. Since drawing ability probably correlates very poorly with either reading ability or general intelligence, the Burgess Scale seriously violates one of its own principles. The subject matter of the scale may be criticized as artificial and unreal, since description of something to do with pictures is not what the child ordinarily reads about. Despite these criticisms the scale has eliminated some of the serious defects of most scales and

is probably one of the best measures of silent reading yet devised.

The Courtis Silent Reading Test

Description of the Test—Part I of this test consists of a story of two ordinary pages in length. The pupil is directed to read silently and is given three minutes to read as much as he can. He then proceeds to Part II which consists of the same story that has just been read but here it is broken up into short paragraphs. A series of five questions to be answered by "yes" or "no" follow each paragraph. There is a preliminary paragraph as an example, after which the pupil is given five minutes to answer as many questions as he can, in order. He is allowed to refer to the paragraphs for the answers. The test is constructed for use with grades 2 to 6.

Scoring the Test—The comprehension score is the number of questions answered correctly minus the number answered incorrectly, divided by number answered correctly. For example, if the pupil answers 28 questions correctly but gives a wrong answer to 12 other questions, his score is 16 divided by 28 or 57%. This score is called the Index of Comprehension. The norms for the different school grades are as follows:

THE COURTIS SILENT READING TEST

School Grade	2	3	4	5	6
Index of Comprehension score	59	78	89	93	95

Norms are also given for the number of words read per minute. This rate is here presented in relation to the norms given by Starch and W. S. Gray.

WORDS READ PER MINUTE

School Grade...	2	3	4	5	6	7	8
Starch	108	126	144	168	192	216	210
Gray	90	138	180	204	216	228	240
Courtis	84	113	145	168	191		

The differences in these norms are no doubt partly to be accounted for by the difference in the subject matter read.

Function of the Test—The subject matter of the Courtis Test is unusually interesting, a factor that is lacking in many reading tests. The tests are easily scored as all answers are either right or wrong. Some criticism may be made to questions answered by "yes" or "no." The effect of guessing is, of course, eliminated by the method of scoring. If all the answers were guessed, half the answers would be right and half wrong and the difference would be zero. It seems that the method of scoring is likely to penalize the pupil who works rapidly, as there is more opportunity for error when more work is done. For this reason the child who works very slowly but carefully, will receive as high a score or likely a higher score on comprehension than the rapid worker. The test has the advantage that rate of reading is measured in half minute periods and comprehension in minute periods. This gives the teacher insight into what type of worker the pupil is.

The Gray Silent Reading Test

Description of the Test—The reading material for this test consists of three short stories. One story, "Tiny Tad," is for use in the second and third grades; another,

"The Grasshopper," is for the fourth, fifth, and sixth grades; the other story, "Ancient Ships," is for use in the seventh and eighth grades. This is an individual test and both rate and comprehension are scored.

Scoring the Test—Rate of reading is measured by the time required to read a page in the middle of the story. The comprehension score is the average of two scores. One is derived from the number of words used in the reproduction of the story by the pupil and the other upon his answers to a list of questions based upon the story. Standards are given for both comprehension and rate.

THE GRAY SILENT READING TEST

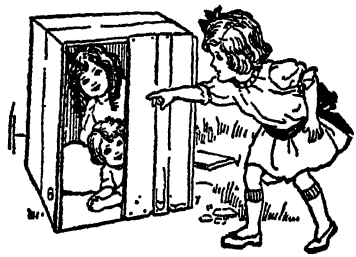
School Grade.....	2	3	4	5	6	7	8
Quality score.....	32	37	29	32	39	22	27
Words read per second	1.50	2.30	2.20	2.57	2.74	2.69	2.87

Function of the Test—This test gives one of the best methods of measuring reproduction that we have, better than either reproduction or questions alone. Since rate of reading is also measured, the test is diagnostic. By its use the pupils' weaknesses in reading may be discovered. The test has the serious practical disadvantage of being an individual test and takes too much of the teacher's time to give it. Furthermore, the method of scoring is difficult and depends too much upon the judgment of the teacher.

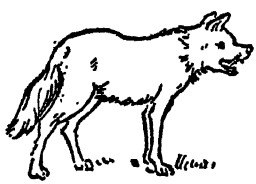
Haggerty Reading Examination

Description of the Test—The Haggerty Examination consists of three parts. Part One, called Sigma 1, is for grades 1 to 3. Sigma 2 is for grades 3 to 6

8. Put a cross over each bird that is on the ground.



9. Put a cross on each child that is hiding.
10. Put two lines under the girl who has found the children.



Once a hungry wolf was about to eat a poor little pig.
The little pig jumped into a big kettle and saved herself just in time.

11. Put a line under the animal which was about to eat the pig.
12. Put a cross under the place where the pig hid.

[3]

FIG. 11. A PAGE FROM THE HAGGERTY READING EXAMINATION—SIGMA 1—FOR GRADES 1 TO 3

and Sigma 3 for grades 5 to 12. Sigma 1 is made up of a preliminary exercise and two tests. Test 1 contains eight sets of pictures, several paragraphs mostly descriptive of the pictures and twenty-five problems based on the pictures and paragraphs. The score on this test is the number of items answered correctly. Test 2 contains twenty questions each followed by "no," "yes," the correct answer to be underlined. Twenty minutes is allowed for the examination.

Sigma 2 is not yet completed. Sigma 3 is composed of three preliminary tests and three regular tests. Test I is a vocabulary test consisting of fifty words, each followed by four or more words, one of which is a synonym or definition. The pupil is to underline the best definition for as many of the fifty words as he can. Test 2 on sentence meaning consists of forty statements followed by "yes," "no." Test 3 on paragraph reading is composed of a series of seven paragraphs each followed by true and false statements. Twenty minutes is allowed for this examination.

Method of Scoring Papers—The score for all these tests is the number of questions answered correctly. In Sigma 1 test 1 and test 2 are scored separately. Provisional norms based upon the results for 6,000 children are given as follows:

NORMS FOR THE HAGGERTY READING EXAMINATION—SIGMA 1

School Grade	1	2	3	4
Score Test 1	4	12	16	20
" Test 2	2	8	14	18

In Sigma 3 the scores for all three tests are combined. Provisional norms are given as follows:

NORMS FOR THE HAGGERTY READING EXAMINATION—SIGMA 3

School Grade....	5	6	7	8	9	10	11	12
Score	21	50	68	76	84	90	96	102

Function of the Examination—In general the Haggerty Examination has much to commend it as a measure of silent reading ability. It measures word meaning, sentence meaning and paragraph meaning separately. It is, therefore, a diagnostic test and the teacher can well use it to discover the cause of poor silent reading. There is little writing required of the pupil and yet the material is not artificial.

Materials Needed

Burgess, May Ayres, A Scale for Measuring Ability in Silent Reading for grades 3 to 8. Four equivalent forms. Russell Sage Foundation, New York City. Sample copy 5 cents, $1.25 per hundred.

Courtis, S. A., Silent Reading Test for grades 2 to 6. S. A. Courtis, 1807 East Grand Blvd., Detroit, Mich. $1.25 for material for testing forty children.

Gray, Wm. S., Oral Reading Test for grades 1 to 8. Public School Publishing Co., Bloomington, Ill. Sample set 6 cents, $1.00 per hundred.

Gray, Wm. S., Silent Reading Tests, test I for grades 2 and 3, test II for grades 4, 5, and 6, and test III for grades 7 and 8. Department of Education, University of Chicago, Chicago, Ill.

Haggerty, J. E., Reading Examinations, Sigma I for grades 1 to 3, Sigma III form A and B for grades 6 to 12, World Book Co., Yonkers, N. Y. Complete material for testing twenty-five children, $1.40.

Monroe, W. S., Standardized Silent Reading Tests, Revised. Test I for grades 3, 4, and 5, Test II for grades 6, 7, and 8, three forms of each. Public School Publishing Co., Bloomington, Ill. Sample set 10 cents, 30 cents per hundred.

Thorndike, E. L., Visual Vocabulary Scales, for grades 3 to 10. Scales A2x, A2y, Bx, and By of approximately equal difficulty. Bureau of Publications, Teachers College, Columbia University, New York City. Manual of Directions, 40 cents, stencil 5 cents, Scale 50 cents per hundred.

Thorndike-McCall Reading Scale, for grades 2 to 12. Ten equivalent forms. Bureau of Publications, Teachers College, Columbia University, New York City. All materials needed supplied with the order, $2.00 per hundred.

Pressey, S. L. and L. C., First grade Reading Vocabulary Test. O. R. Chambers, 215 East Third Street, Bloomington, Ind. Price per package of one hundred, all materials included, 60 cents.

Selected References

Gray, C. T., *Types of Reading Ability as Exhibited through Tests and Laboratory Experiments,* School of Education Monographs, University of Chicago, Chicago, Ill.

Gray, W. S., *Studies of Elementary School Reading through Standardized Tests,* School of Education Monographs, University of Chicago, Chicago, Ill.

Huey, E. B., *The Psychology and Pedagogy of Reading,* Macmillan.

Judd, C. H., *Reading, Its Nature and Development,* School of Education Monograph, University of Chicago, Chicago, Ill.

O'Brien, J. A., *Silent Reading,* Macmillan.

Stone, C. R., *Silent and Oral Reading,* Houghton Mifflin Co.

Smith, W. A., *The Reading Process,* Macmillan.

CHAPTER VIII

ENGLISH LANGUAGE AND COMPOSITION

Just as ability to read is fundamental to excellence in all school subjects, so the ability to express one's thoughts, ideas and feelings is fundamental to the higher levels of school attainment. But correct expression of thought is a difficult object for measurement, because of the complexity of the factors involved. Judges of language are continually finding themselves in controversy over proper standards for judgment of both the written and the spoken word.

General recognition is made of the fact that the ideal speech or essay must contain certain elements, to be considered satisfactory, but the weight or value to be assigned to each of these is by no means determined. Thus it is pretty well conceded that a good composition must conform to certain generally accepted standards of mechanical detail, as in matters of good spelling, correct punctuation, legible penmanship; to standards of structural form, as sentence structure, paragraph building, felicity of diction and choice of expression; and standards of thought content, whereby the writer, through mechanics and form, presents an idea or mental conception of more or less worth. But the relative weights to be allotted to mechanics, to form, and to content, are by no means universally agreed upon by teachers or critics. Accordingly the teacher is quite uncertain as to proper standards

to be employed in estimating the value of compositions, and in holding a proper balance between the various elements entering into the child's written performance.

The aids given by objective tests and scales are of a varied character, and should all be used by the teacher of composition; they involve the factors underlying good composition, and the finished product as well. Under the first class, are those scales which have to do with the choice and meaning of words, the tests which determine ability in punctuation, in grammar, in copying accurately; under the second group will be found tests evaluating specimens of completed composition work by various methods of comparison and classification.

Some of the useful tests in composition have an equally high value in determining some of the elements of the reading process, and have already been described. Under this head are the various vocabulary tests, the use of which reveals some valuable knowledge regarding the child's comprehension of word meanings, and his range of vocabulary. In like manner, the spelling tests, which are of course of assistance in diagnosing difficulties in this important mechanical side of composition work, have been already described in the appropriate chapter.

Punctuation Scales

Several scales have been designed to test punctuation ability. The first to have wide use was the Starch Punctuation Scale.

The Starch Punctuation Scale

Description of the Scale—These scales are made up of a series of exercises arranged in "steps," each exercise

consisting of sentences to be punctuated by the pupils. Each "step" of the scale is made up of sentences so selected that the difference in difficulty between any two successive steps of the scale is equal to the difference between any other two successive steps. Sample "steps" are as follows:

STEP 7

1. I told him but he would not listen.
2. Concerning the election there is one fact of much importance.
3. The guests having departed we closed the door.
4. The train moved swiftly but Turner arrived too late.

STEP 11

1. Paris Illinois is a smaller city than Paris France.
2. He asked what is the matter.
3. I like to work he said especially in the morning.
4. Chicago Illinois is a large city.

Use of the Scale—The pupil is handed a printed copy of the scale and instructed to correct the sentences, by inserting the appropriate marks of punctuation where needed. The procedure is decidedly easy.

Function of the Scale—The scale measures ability of the child to distinguish correct forms and usage, and the assumption is that if he can do this accurately in the scale, he will be able to apply the same knowledge in his own written work. On the other hand, his failures in the scale can be easily classified, and form the basis for intelligent drill. Since the application of the scale is not at all difficult, and the scoring easy and therefore likely to be accurate, the use for class information is made simple and therefore to be welcomed by the teacher. The value for diagnosis of this sort of mechanical difficulty for

both elementary and high school pupils is immediately evident. And this brings out one fault in the nature of the scale. The sentences used do not bring out in sufficient variety the types requiring variation in punctuation, so that the diagnostic value is not as great as though the entire matter of punctuation usage were more carefully analyzed, and covered in the scale. Even with this fault, it remains the most satisfactory of the various punctuation scales.

English Grammar Scales

Starch's English Grammar Scales

Description of the Scale—The Starch Scales consist of three distinct series of exercises, constructed in the same general way as his punctuation scales, each set being arranged in a series of "steps." Each step consists of a group of exercises, requiring knowledge on the part of the pupil of certain definite language forms. The idea in relation to language usage, is to determine whether the pupil can use the forms correctly, rather than to test his knowledge of the rules on which they are based. Typical exercises are as follows:

Step 8. (*Scale* B)

1. The fact that I had never before studied at home, (I was at a loss; made me feel at a loss as to) what to do with vacant periods.
2. Both are going,—(he and she; him and her.)
3. I do n't believe I (will; shall) be able to go.
4. It is (the handsomest vase I almost; almost the handsomest vase I) ever saw.

STEP 11

1. There we landed, and having eaten our lunch (the steamboat departed; we saw the steamboat depart).
2. (After pointing; when he had pointed) out my errors, I was dismissed.
3. The question of (whom; who) should be leader arose.
4. He spoke to some of us,—namely (she and I; her and me.)

Method of Using the Scale—Printed copies of the scale are given to the students, and they are instructed: "Each of the following sentences gives in parenthesis two ways in which it may be stated. Cross out the one you think is incorrect or bad. If you think both are incorrect, cross both out. If you think both are correct, underline both." The score is based on the highest step in which the pupil does seventy-five per cent of the exercises. Starch suggests the following scores as standards of attainment:

Grade...	7	8	9	10	11	12	Freshmen
Score ..	8.0	8.3	8.6	8.9	9.2	9.5	10.3

Function of the Scale—As has been said, the scale is designed to test ability to use grammatical forms, not to test knowledge of rules of grammar, or terminology of the subject. It carries out the idea that if the pupil has a sufficient knowledge of usage, the teacher need not attempt formal instruction in grammar. A large number of grammatical forms are included in the scale, and the instructor is therefore given ample opportunity to determine weaknesses of the pupils. The scale itself has been criticized because it is not strictly diagnostic, in that the various forms are not arranged in a systematic manner, so as to make easy the recognition of types of weak-

nesses. However, the properly informed teacher will not find it difficult to make her own diagnosis for each pupil.

The Charters' Diagnostic Language Tests

Description of the Scale—Dr. Charters has designed scales to test the use of pronouns and verbs in context so that the test also involves knowledge of good language usage. Form 1 involves pronouns and is designed for use in grades 3 to 12 inclusive. It is paralleled by a Form 2 for alternate use. There are two forms also for the verb test, but this is designed for use in the higher grades only. In the pronoun test forty sentences are given in which pronouns are used, correctly in some sentences, incorrectly in a majority. The pupil is to indicate on the printed form whether the usage is correct or not, and if the latter, the proper usage. Sample sentences are:

7. Who do you want?
14. It was only us.
33. Who did you speak to?

Method of Using the Test—Printed blanks are given to the pupils. They are then instructed as follows: "This test is given to pupils who have studied language lessons to see how well they are able to tell when sentences are right and when they are wrong. Now look at the sample below:

I told him to go.

.....

"The plan is to read this sentence over carefully and see if it is right. If it is right, make a cross on the dotted

line below the sentence. The sentence 'I told him to go' is right, so we shall make a cross on the dotted line below it. Make the cross now.

"If the sentence is not right, we are to put the correct word or words on the dotted line below it. Let us try one that is not right, etc." After this careful instruction, the pupils proceed to work out the paper, being given ample time to finish, as the element of speed does not enter. A scoring key accompanies each form of the test, and careful directions are given for recording scores properly. Each scoring key is accompanied also by standards suggested by the author.

Function of the Scale—The scale as described measures the ability to use correct forms of pronouns and verbs. Dr. Charters has also worked out one form in which the pupil is given an opportunity to give reasons for corrections, and this tests the knowledge of the rule or principle governing the proper use of the form. As is indicated by the name, the teacher is expected to diagnose the pupil's difficulty, and to supplement this by proper instruction and drill. The author bases the tests upon his researches in language errors. Thus, he collected twenty-five thousand errors made by pupils in using pronouns in oral speech. Upon study, it developed that there were only forty different kinds of errors involved in the entire twenty-five thousand total. The test is designed to bring in as many of these forty errors as possible, and so to enable the teacher to determine those most likely to occur in her class, and so to determine the direction of her drill. The scientific nature of the underlying basis of the test makes the resulting diagnosis a most useful one in laying the proper foundation work for usage in the matter of

pronouns and verbs, and also brings out incidentally a number of miscellaneous weaknesses in usage.

Tests of Diction

The Trabue Completion—Language Scales

Description of the Tests—Dr. Trabue has worked out on a scientific basis a considerable number of scales, lettered from A to M, and also a Scale Alpha and a Scale Beta. There have also been some adaptations of his scales by other experimenters. In the main, they follow one general form. A sentence is given, in which one or more words are left blank, and the pupil is required to fill in the missing word, thus completing the sentence. Scales B, C, D, E, and F, are of equal value, and so may be used interchangeably to avoid danger of "coaching." Sample sentences from Scale D are:

1. We are going.....school.
5. Hard.....makes.....tired.
8. The best advice.....usually.....obtained .....one's parents.
10. .....a rule.....associations..... friends.

Method of Using the Scale—Printed blanks are distributed to the pupils, and they are instructed to write one word in each blank, in each case writing that word which makes the best sense. Sample sentences are given for practice by the children, so that the method of procedure may be entirely clear. A time limit is imposed for each scale, which must be carefully observed. A scoring key is furnished with the scales, which makes the de-

termination of right and wrong usage very clear. Allowance is made for slight imperfections, so that half credit is given when the sentence is less than perfect, but still makes reasonably good sense.

Function of the Scale—Trabue describes the scale as "an attempt to derive one or more scales for the measurement of ability along certain lines closely related to language." Other psychologists have described it as a test of "general ability" or of "intelligence." Certainly the scale involves a knowledge of grammar, an ability to reason, to determine the fitness of various words in their relations, and so a knowledge of diction and some range of vocabulary. Norms have been worked out for all grades of the grammar school and the four years of the high school. The scale is described as "for children between seven and twenty years old." On account of its high correlation with tests of "general intelligence," the scale may serve a double purpose, either to indicate the intellectual level of the individual or the class, or to determine language development. As a test of language, it may be used as prognostic, or as diagnostic, to determine corrective measures for the pupil or the class; or it may be used as an achievement test in language, to find out the relative development of the language ability of the individual or group in question.

Tests of Composition

Up to this point tests and scales have been considered which are to be placed in the hands of the pupils, in order to discover their ability along lines involved in language work, but which do not measure the completed composition. A series of scales have been devised which give

standards for judging completed paragraphs or entire compositions on a basis of general quality, or of elements comprehended in the expression "general quality." These scales are made up of selected paragraphs indicating various grades of composition attainment, and the teacher is expected to compare the actual work of the pupil with the selections on the scale, and to assign a value to the composition equal to that of the scale paragraph which it most nearly approaches in value. This involves a certain question of judgment, and for this reason these scales are not so nearly objective as are those hitherto described in this chapter. Accordingly, teachers have been divided in their opinions as to the value of such scales. In general it may be said that those teachers who have practised the use of the composition scales have found them a great help in setting reasonable standards, and so have been able to arrive at judgments on pupils' work much more accurate than those arrived at by purely subjective methods. The first use of such a scale is likely to be relatively inaccurate; but practice makes it a valuable adjunct to the teacher in both elementary and high schools.

The Nassau County Supplement to the Hillegas Scale

Dr. Milo B. Hillegas was the pioneer in devising a scale for measuring English composition, and his scale, published in 1912, has been made the basis for several succeeding scales. The general method of making the scale was to select a series of compositions, most of them actual work of school children, arranged in order of merit, as determined by consensus of a large number of competent judges, and corrected by careful statistical methods. The

original scale was not entirely satisfactory since the selections were of differing length and character, no two dealing with the same theme, and separated by irregular intervals. Several attempts have been made to correct these difficulties; Prof. E. L. Thorndike has made an extension of the scale, and Prof. M. R. Trabue devised the Nassau County Supplement.

Description of the Scale—The scale is intended for use in grades four to twelve inclusive. It consists of ten samples of composition ranging in value from 0 to 9.0, arranged in ascending order on one sheet. The actual values are: 0, 1.1, 1.9, 2.8, 3.8, 5.0, 6.0, 7.2, 8.0, 9.0. Thus they approximate closely a ten unit division of a scale. The first seven samples are compositions written on the subject "What I should like to do next Saturday," and were obtained in a survey of the schools of Nassau County, N. Y. Specimen 10 is taken from literature. The entire scale is intended to measure "general quality," and there is no indication of the factors involved in this determination.

Method of Using the Scale—The author gives directions for obtaining compositions for comparisons, suggesting that standard conditions be observed. Thus, the topic assigned "must be interesting and suggestive to the pupils, and at least twenty minutes must be allowed for the writing. 'What I should like to do next Saturday' will produce a higher average quality of results than 'How to Play Baseball,' but it will probably produce a lower average than 'The Most Exciting Experience of My Life.' "

The teacher is then told to compare the general quality of the composition with the general qualities of the vari-

ous samples on the scale, and to assign to the composition the numerical value of the printed sample which it most nearly equals in general merit.

Function of the Scale—As has been indicated, the whole idea of the scale is to measure "general quality" of English composition. As there is no attempt to analyze general quality, it is valuable rather as a measure of general attainment of the pupil or the class, than as a diagnostic or prognostic measure. Teachers untrained in the use of the scale differ widely upon the first application of it, but experiments have demonstrated that after a few hours of training, the applications become more objective, and variation between teachers becomes relatively slight. As a general measure it has certainly a definite value.

Other scales—Other scales arranged on the same general plan as the Nassau and Hillegas Scales are the Breed and Frostic, especially adapted to the sixth grade; the Willing, for grades four to eight, which involves the factors of story value, spelling, punctuation, capitalization and grammar, and the Hudelson, which is a supplement to the Hillegas Scale based upon one thousand compositions written by first year high school pupils in Virginia. The Hudelson Scale gives a more uniform value from step to step than does the Nassau Supplement, and is accompanied by complete directions for scoring compositions which make a valuable addition to the language teacher's equipment.

A somewhat different type of scale is represented by the Harvard-Newton Scale, devised under the direction of F. W. Ballou and arranged in four parts, one for each form of discourse, narration, description, exposition, and

argumentation. Each of these parts forms a distinct scale, and so makes a valuable instrument where the composition work of a school is taught in accordance with these divisions. The idea involved in the Harvard-Newton Scale is also used in the Minnesota Composition Scale, devised by Dr. M. J. Van Wagenen. This scale is arranged in three parts, a separate scale for each of the forms of discourse, narration, description and exposition. The scales are made up of either fourteen or fifteen selections, actual compositions written by Minnesota school children. Each scale contains compositions written upon the same general topic, that in Narration being "When Mother was away," in Description, "It was a Sight worth seeing when the Troops marched by," and for Exposition, "How I earned some Money." Each sample is given a definite value on each of three factors, "structure," "mechanics" and "thought content." In this way, it becomes possible to compare and evaluate work done along three distinct lines, thus making a sort of diagnosis possible which is not achieved in the other scales described.

Mention should also be made of Lewis's Special-Type Scales, two of which are devised to measure ability to write business letters, two to measure achievement in writing friendly letters, and one for general composition, arranged in somewhat the same way as the Harvard-Newton Scales.

When the entire field is surveyed, there is certainly every ground for believing that a teacher who has familiarized herself with the various sorts of measuring instruments suggested in this chapter will be a much more able teacher of language than if she had not gone into the

study of structure, mechanics and content involved in the use of these scales.

Materials Needed

Ballou, F. W., Harvard-Newton Composition Scale. Harvard University Press, Cambridge, Mass.

Charters, W. W., Diagnostic Language Test, for grades 3 to 8. Public School Publishing Co., Bloomington, Ill. Sample set 10 cents, 80 cents per hundred.

Lewis, E. E., Scales for Measuring Special Types of English Composition, for grades 3 to 12, World Book Co., Yonkers, N. Y. Monograph including directions for giving and scoring, $1.36.

Starch, D., Punctuation Scale. University Coöperative Store, Madison, Wis.

Starch, D., English Grammar Scale, University Coöperative Store, Madison, Wis.

Trabue, M. R., Completion Test Language Scales, for grades 2 to 12 and above. Eleven different forms, B, C, D. E and F are equivalent, L and M are equivalent and especially intended for use in high school. J and K are equivalent and for use with college students and adults. Alpha and Beta are equivalent and are combinations of the other tests. A Key to the Trabue Language Scales (A Manual) 75 cents. Key for forms B and C 20 cents. Sample copy of each scale, 10 cents. Scales Alpha and Beta $1.25 per hundred, all others 50 cents per hundred. Bureau of Publications, Teachers College, Columbia University, New York City.

Trabue, M. R., Nassau County Supplement, Teachers College, Columbia University, New York City.

Van Wagenen, M. J., Minnesota Composition Scale, M. J. Van Wagenen, University of Minnesota, Minneapolis, Minn.

Selected References

Ashbaugh, E. J., "The Measurement of Language," *Journal of Educational Research*, June, 1921.

Hillegas, M. B., "Hillegas Scale for Measurement of English Composition," *Teachers College Record*, September, 1912.

Trabue, M. R., "The Nassau County Supplement to the Hillegas Scale," *Teachers College Record,* January, 1917.

Twenty-Second Yearbook of the National Society for the Study of Education, English Composition, Its Aims, Methods, and Measurements, Public School Publishing Co., Bloomington, Ill.

CHAPTER IX

ARITHMETIC

Arithmetic is probably the next most important school subject after reading. Its principles are, in general, fundamental to success in school and in business. While emphasis is often misplaced in the teaching of arithmetic, it no doubt rightly retains its important place in determining the promotion or failure of a pupil. The three R's have been and no doubt will long remain the fundamental part of the work of the first few years of school. What has been said applies to the fundamentals of arithmetic as taught in the grades and not to high school or college mathematics. The relative importance of higher mathematics is another problem that does not warrant consideration at this time.

The first number concepts develop very early in the child's life. The child soon learns to distinguish between self and not self. The persons and things of the environment are also early classified as few or many. Very likely at first the child makes a distinction only between one and two as contrasted with larger numbers. Slowly but gradually this concept of numbers and quantity grows and becomes more definite. That this growth was slow in some races, at least, is evidenced by the fact that the American Indians characterized any number more than ten as a "heap big many." It is interesting in this connection to point out the relationship between the number of fingers and ten as the basis of our decimal system.

This shows that the fingers have had a large part in the development of the number concepts of the race as well as of most children.

When the fingers or anything else is used as the basis of counting the process becomes one of abstraction. In fact, as soon as the child begins to count objects it begins to abstract the property of number from other properties, such as size, weight, color, etc. This process in its simpler forms commences long before the school age. In school this developing capacity is made the basis of the higher and more complex concepts of numbers. The pupil is confronted by the symbolism of numbers in arithmetic as well as by the symbolism of reading.

This symbolism of numbers is soon further complicated by the processes of the fundamental operations of addition, subtraction, multiplication and division. The pupil must learn to manipulate these abstract symbols rapidly and accurately. All this precedes the application of these processes to the solving of problems. But problem solving involves more than the manipulation of numbers. The pupil must determine what numbers are to be used in the solution and what processes are to be used with the numbers. Problem solving, therefore, introduces another and for the most part a higher form of reasoning for the child.

In the construction of tests a distinction has been made between the fundamental operations and problem solving. In the former we have such tests as the Courtis Standard Research Tests, The Cleveland Survey Tests and the Woody-McCall Mixed Fundamentals. In the latter there are such tests as the Monroe Reasoning Tests and the Rogers Diagnostic Tests.

The Courtis Standard Research Tests—Series B

Description of the Tests—These tests consist of a series of problems in each of the four fundamental operations. There are twenty-four problems in each operation. In Addition each problem contains nine three place numbers. The problems are of equal difficulty and the class is allowed eight minutes to solve as many problems as possible. Four minutes is allowed for the problems in Subtraction, six for the problems in Multiplication and eight for Division. The problems of each of the four operations are arranged on separate pages of the test folder.

Scoring the Tests—Score cards are provided to facilitate the scoring of the papers. Only those problems with correct answers receive credit but the number attempted is also recorded as a basis for determining accuracy. The scores are recorded on a class record sheet which is furnished with the test. Complete directions are furnished for computing median scores and median differences. The method is somewhat complicated but carefully described so that the classroom teacher may follow the directions step by step. Tentative standards are given as follows for June.

STANDARDS FOR COURTIS STANDARD RESEARCH TESTS

GRADE	*Test 1* ADDITION	*Test 2* SUBTRACTION	*Test 3* MULTIPLICATION	*Test 4* DIVISION
3.....	4	5	0	0
4.....	6	7	6	4
5.....	8	9	8	6
6.....	10	11	9	8
7.....	11	12	10	10
8.....	12	13	11	11

Sample Problems from the Courtis Standard Research Tests

Eight of the Twenty-Four Problems in Addition

You will be given eight minutes to find the answers to as many of these addition examples as possible. Write the answers on this paper directly underneath the examples. You are not expected to be able to do them all. You will be marked for both speed and accuracy, but it is more important to have your answers right than to try a great many examples.

927	297	136	486	384	176	277	837
379	925	340	765	477	783	445	882
756	473	988	524	881	697	682	959
837	983	386	140	266	200	594	603
924	315	353	812	679	366	481	118
110	661	904	466	241	851	778	781
854	794	547	355	796	535	849	756
965	177	192	834	850	323	157	222
344	124	439	567	733	229	953	525

Eight of the Twenty-Four Problems in Subtraction

92971900 62207032	104339409 74835938	60472960 50196521	119811864 34379846
137769153 70176835	144694835 74199225	123822790 40568814	80836465 49178036

Ten of the Twenty-Five Problems in Multiplication

6385 48	8736 502	5942 39	6837 680	4952 47
3876 93	9245 86	7368 74	2594 25	6495 19

Eight of the Twenty-Four Problems in Division

$25\overline{)6775}$	$94\overline{)85352}$	$37\overline{)9990}$	$86\overline{)80066}$
$73\overline{)58765}$	$49\overline{)31409}$	$68\overline{)43520}$	$52\overline{)44252}$

Function of the Tests—The Courtis Tests are among the best known of the educational tests. Arithmetic lends itself to exact measurement and hence it was one of the first of the school subjects for which tests were constructed. The tests provide an easy means whereby the teacher may compare her class with what should be expected for that grade. They also provide a very practical teaching device for drill in the fundamental operations without disturbing the processes of instruction for the other children in the grade.

The Courtis Standard Practice Tests in Arithmetic

Description of the Tests—These are not ordinary tests but a series of carefully graded problems for daily practice. The material for practice consists of two parts: (1) a set of Lesson Cards and (2) a Student's Record and Practice Pad. The lesson cards are series of problems in addition, subtraction, multiplication and division. The first card, for example, contains 72 simple problems in addition. The answers to the problems are printed on the back of the card. The Practice Pad contains sheets of transparent paper and one of the lesson cards is inserted under a sheet of paper so that the pupil may work the problem on the transparent paper.

On the first day a preliminary test is given and all children reaching a certain standard are excused from drill. Those pupils who need the drill spend from three to four minutes each day in practice on each lesson card—the amount of time spent in practice depends upon the school grade. The pupil is trained to score his own work by comparing his answers with those given on the back of the cards. The score made each day is recorded and

Sample Lesson of the Courtis Standard Practice Tests

Lesson No.....Form.....Date.....

Name..... Grade.....

21 14 —	81 12 —	31 16 —	71 13 —	51 17 —
61 15 —	41 18 —	51 19 —	71 52 —	31 27 —
41 24 —	51 45 —	61 26 —	71 69 —	81 23 —
51 38 —	22 22 —	32 31 —	42 23 —	52 34 —
62 32 —	72 33 —	32 34 —	52 44 —	73 23 —

5 10 15 20 25

10 B

SCORES	COURTIS STANDARD	TRIAL
Tried —	PRACTICE TESTS	No.——
Right —	LESSON No. 10 MULTIPLICATION	Form B

graphs are drawn by the child so that the pupil may see the effect of his own practice. Practice is continued on any card until standard efficiency is reached. The pupil then passes on to the next practice card. This necessitates individual instruction in arithmetic but the practice cards make this possible. Each pupil may progress at his own rate. There are 48 cards containing lists of problems arranged by gradual steps for daily practice. The pupil can be drilled on these problems, and tests are provided from time to time for measuring progress.

The Teacher's Manual describes in detail methods for teaching each of the simple operations in arithmetic. There are two forms, A and B, of the Standard Practice Tests of equal difficulty and, therefore, interchangeable.

A recent criticism [1] has been made of the practice tests that the arrangement of the number combinations gives an equal amount of practice on all the combinations. Some combinations are much more difficult than others for the child and, therefore, more practice should be provided for the more difficult combinations and less for the easier ones. For example, 8 and 9 need more practice than combinations of 1 and 2. This criticism seems valid and no doubt the material of the tests should be revised on the basis of the difficulty of the number combinations.

The Studebaker Economy Practice Exercises

The Studebaker Economy Practice Exercises published by Scott, Foresman & Co. are so similar to the Courtis Practice Tests that no separate description of these tests

[1] W. J. Osburn, "A Study of the Validity of the Courtis and Studebaker Tests in the Fundamentals of Arithmetic," *Journal of Educational Research,* Vol. VIII, No. 2, September, 1923.

will be given. They are based upon the same general principles, serve the same purpose and are open to the same criticism as the better known Courtis Tests.

The Cleveland Survey Arithmetic Tests

Description of the Tests—These tests consist of fifteen sets of exercises in the four fundamentals of arithmetic, and fractions.

The sets are arranged in spiral form, that is, the same operations recur several times in the test but each time the problems introduce some new factor or factors. The first problems in addition consist of a set of two single place numbers. These are followed by sets of simple problems in subtraction, multiplication, and division, then in turn by a set of problems consisting of five single place numbers to be added. This spiral arrangement is followed throughout the test, each new set of problems being more difficult than the preceding. The eighth and fifteenth sets are problems in the addition, subtraction, multiplication, and division of fractions.

Method of Giving and Scoring the Test—The pupils are given from 30 seconds to 4 minutes, differing for the various tests, in which they are to work as many problems in each set as they can. The total actual working time for the fifteen sets of problems is 22 minutes. The problems are scored by means of a key. The number of problems right in each set constitutes the final score. There are no general standards or norms for these tests. Standards for St. Louis and Grand Rapids are given on the folder accompanying the tests. There is some variability between these scores; the St. Louis scores in general are the higher.

SAMPLE PAGE FROM THE CLEVELAND SURVEY ARITHMETIC TESTS

						Ats.	Rts.
SET L—Multiplication—							
8246		3597		5739	2648		
29		73		85	46		
4268		7593		6428	8563		
37		64		58	207		
SET M—Addition—							
7493	8937	8625	2123	5142	3691		
9016	6345	4091	1679	0376	4526		
6487	2783	3844	5555	4955	7479		
7591	4883	8697	6331	9314	2087		
6166	1341	7314	6808	5507	8165		
5226	9149	6268	9397	7337	8243		
2883	8467	7725	6158	2674	6429		
2584	0251	8331	3732	9669	9298		
0058	7535	5493	4641	5114	7404		
2398	5223	3918	7919	8154	2575		
SET N—Division—							
$67\overline{)32763}$		$48\overline{)28464}$		$97\overline{)36084}$	$59\overline{)29382}$		
$78\overline{)69888}$		$88\overline{)34496}$		$69\overline{)40296}$	$38\overline{)26562}$		

Function of the Tests—The Cleveland Survey Tests are highly diagnostic, much more so than the Courtis Tests, since they analyze the fundamental processes into suc-

STANDARDS (MEDIAN NUMBER OF EXAMPLES CORRECT) FOR THE CLEVELAND SURVEY ARITHMETIC TESTS

ST. LOUIS, MISSOURI

Grades

TEST	3–B	3–A	4–B	4–A	5–B	5–A	6–B	6–A	7–B	7–A	8–B	8–A
A	14.6	18.3	19.8	21.3	22.5	22.5	26.3	26.4	27.8	28.4	32.3	32.2
B	9.9	12.2	17.1	17.0	18.0	20.0	20.3	20.6	22.8	24.2	26.7	28.3
C	7.6	10.5	16.7	15.4	16.9	16.7	18.2	18.3	18.9	19.8	20.7	21.9
D	9.0	12.2	15.8	16.3	18.4	17.8	19.3	20.5	21.3	22.3	23.8	25.7
E	3.8	4.8	5.7	5.4	6.0	6.1	6.9	7.1	6.6	7.4	8.0	8.4
F	2.3	3.5	5.6	6.0	6.4	7.4	8.0	8.3	8.5	9.6	10.1	11.3
G	2.7	3.5	4.9	5.1	5.5	5.6	5.9	6.2	6.4	6.9	7.4	7.8
H	0.7	3.8	7.8	6.8	4.8	6.5	8.0	8.1	9.5	9.7	10.8	12.0
I	1.1	1.4	2.0	2.0	3.0	3.2	3.9	4.1	4.5	5.0	5.4	5.8
J	1.6	2.9	3.8	3.9	4.1	4.3	5.0	5.1	5.2	5.3	5.4	5.8
K	....	....	3.3	4.0	5.0	5.8	6.9	7.4	8.3	9.7	10.3	11.7
L	....	....	2.5	2.9	3.1	3.4	4.3	4.1	4.6	4.7	5.2	5.3
M	0.5	2.1	2.9	3.3	3.4	3.7	4.2	4.4	4.5	4.9	5.2	5.3
N	....	....	0.8	1.1	1.3	1.4	1.6	1.8	2.0	2.0	2.6	2.7
O	....	....	....	2.5	3.3	3.3	3.6	4.1	4.8	5.6	6.1	6.6

GRAND RAPIDS, MICHIGAN

Grades

TEST	3–B	3–A	4–B	4–A	5–B	5–A	6–B	6–A	7–B	7–A	8–B	8–A
A	11.8	13.4	13.6	16.4	20.3	21.5	22.8	25.0	26.5	27.3	29.5	30.3
B	6.3	8.4	9.1	12.1	14.7	15.9	16.8	19.1	21.3	20.7	22.8	25.5
C	....	....	7.1	11.3	13.7	14.0	15.5	17.0	17.7	18.8	19.3	20.7
D	....	....	6.9	10.4	12.5	14.3	15.5	16.9	18.4	19.7	20.5	23.0
E	....	....	4.1	4.6	5.2	5.4	6.0	6.6	7.2	7.2	7.8	8.1
F	....	....	2.8	4.1	6.0	6.5	7.1	8.0	9.3	9.6	10.3	11.0
G	....	....	2.2	3.3	4.5	4.9	5.3	5.6	6.1	6.1	6.7	6.8
H	....	....	....	....	....	6.3	6.2	6.5	9.0	7.8	8.6	8.8
I	....	....	0.7	0.9	1.3	1.4	2.3	3.0	3.8	4.1	4.0	4.7
J	....	....	....	2.8	3.4	3.7	4.1	4.5	5.4	5.3	5.7	6.5
K	....	....	....	....	3.0	4.3	5.4	6.5	7.5	8.8	9.7	10.3
L	....	....	....	....	2.3	2.9	3.3	3.6	4.3	4.5	4.9	4.9
M	....	....	....	2.3	3.0	3.6	4.3	4.5	4.9	5.0	5.7	5.7
N	....	....	....	....	0.7	0.8	1.1	1.4	1.7	1.8	2.0	2.3
O	....	....	....	....	....	....	3.5	3.6	3.9	4.6	5.5	4.8

cessive steps. The teacher may determine what types of problems the pupil can solve in addition, for example, and what types he can not solve. This gives an indication of

the pupil's difficulties. By the use of the tests the teacher can determine where wrong principles are involved, where drill should be placed and where it is not necessary. The tests are not so usable as some of the others for survey purposes, since such norms as are given are for each separate test and not for the whole test. If these were combined norms given for each of the fundamental operations the tests would be more serviceable for survey purposes but less valuable for diagnostic uses.

The Woody-McCall Mixed Fundamentals

Description of the Test—This test consists of 35 problems in addition, subtraction, multiplication and division of whole numbers, common and decimal fractions. The different operations are arranged in irregular order on the test sheet. In some of the problems the operation is indicated by words and in others by signs. The pupil is allowed 20 minutes in which to solve as many problems as possible. The score is the number of problems having correct answers expressed in lowest terms. Norms are given for October scores as follows:

THE WOODY-McCALL STANDARDS

School Grade.....	3	4	5	6	7	8
Standard score.....	6.8	13.1	17.8	22.5	25.9	27.8

In order to make these standards comparable with results obtained later in the school year for each month after October add the following increments:

School Grade.....	3	4	5	6	7	8
Increment to be added for each mo.....	.54	.43	.42	.24	.25	.20

Norms are also given for high and low sections of each class separately.

Showing Twenty of the Thirty-Five Exercises of the Woody-McCall Mixed Fundamentals Test

WOODY-McCALL MIXED FUNDAMENTALS FORM I

Name..... *Age*.....

Grade.....

Get the right answer to as many examples as you can in 20 minutes. Do all work on the front or back of this sheet.

(1) ADD	(2)	(3)	(4) SUBTRACT	(5) MULTIPLY	(6) SUBTRACT
2	$2 \times 3 =$	$3\overline{)6}$	2	23	13
3			1	3	8
—			—	—	—

(7) ADD	(8)	(9) SUBTRACT	(10) MULTIPLY	(11)
17	$3 + 1 =$	16	254	$4 \div 2 =$
2		9	6	
—		—	—	

(12) ADD	(13) SUBTRACT	(14)	(15) ADD	(16) MULTIPLY
23	393	$2\overline{)13}$	9	5096
25	178		24	6
16	—		12	—
—			15	
			19	
			—	

(17)	(18) ADD	(19) MULTIPLY	(20)
$2\frac{3}{4} - 1 =$	$12.50	7898	¼ of 128 =
	16.75	9	
	15.75	—	
	—		

Function of the Test—In many respects this test is not different from the others in the fundamentals of arithmetic. The miscellaneous arrangement of the problems probably more nearly simulates ordinary test conditions. The use of signs in many of the problems emphasizes the importance of a knowledge of their meaning. The score is a total of all problems solved correctly. The test is, therefore, to be used as a general survey test. Only by further analytical study can it be used for diagnostic purposes. This limits the usefulness of the test for classroom purposes.

The Woody Arithmetic Scales are similar to the Woody-McCall tests except that the problems of each operation are printed on separate pages of a folder and norms are given for each of the fundamental operations.

The Monroe Diagnostic Tests in Arithmetic

Description of the Tests—These tests are so similar to the Cleveland Survey Tests that only the chief differences between them will be pointed out. Monroe has twenty-one sets of problems in his tests and they contain more difficult problems than the Cleveland Survey Tests. They are printed on four separate sheets. The fourth grade uses the first two sheets, the fifth grade uses the first three sheets, and the sixth, seventh and eighth grades use all four. The time allowances and method of scoring are similar to the Cleveland Tests. Tentative mid-year norms for the number of examples correct are given as follows:

Monroe Diagnostic Tests in Arithmetic

TENTATIVE STANDARDS—MID-YEAR SCORES
NUMBER OF EXAMPLES CORRECT

TEST No.	GRADE IV	V	VI	VII	VIII
1	8.3	8.5	10.2	12.0	12.7
2	3.0	5.3	7.1	8.0	8.9
3	2.2	2.7	4.0	4.6	5.2
4	1.1	1.3	2.3	3.4	4.0
5	2.3	2.7	3.3	3.4	4.0
6	1.1	1.6	2.6	3.3	4.5
7	2.2	2.8	3.4	3.9	4.3
8	1.2	2.3	3.1	4.0	4.4
9	2.7	5.8	6.5	7.5	8.2
10	1.4	1.9	3.4	3.9	5.4
11	.9	1.1	1.6	2.0	2.3
12		1.4	3.5	4.3	5.4
13		1.6	2.5	3.3	3.7
14		1.9	3.8	5.2	5.1
15		1.4	2.7	3.3	3.3
16		1.9	3.4	5.7	6.1
17			1.6	2.2	2.5
18			8.3	9.5	11.0
19			2.4	3.4	3.5
20			8.5	10.0	11.0
21			1.7	2.2	2.4

The Monroe General Survey Scale in Arithmetic is also similar except that the purpose of this scale is to provide a single score which will be a general measure of the pupils' ability to perform the operations of arithmetic.

Monroe's Standardized Reasoning Tests in Arithmetic

Description of the Tests—These tests consist in a series of practical problems such as are to be found in an ordinary text book in arithmetic. They involve the fundamental operations, fractions, denominate numbers and percentage. They are not time tests, for the pupil is

given sufficient time to solve as many problems as he can. There are three separate tests, one for grades 4 and 5, another for grades 6 and 7 and another for grade 8.

Sample Test II for Grades 6 and 7. Form 1

STANDARDIZED REASONING TEST IN ARITHMETIC

Devised by Walter S. Monroe

6. Four loads of hay are to be put into a barn. The first load weighs 1.125 tons; the second, 1.75 tons; the third, 1.8 tons; the fourth 1.9 tons. Find the weight of the four loads.
   P = 1
   C = 2
7. A baker used 3/5 lb. of flour to a loaf of bread. How many loaves could he make from a barrel (196 lbs.) of flour?
   P = 3
   C = 2

The problems are to be scored for both correct principle and correct answer. For example, if a mistake is made in an addition, but the operation is the correct one, a score is given for the correct principle. A key is provided as a guide in scoring the papers. Tentative norms are given as follows:

School Grade	4	5	6	7	8
Correct principle	9.6	17.0	15.5	20.7	16.8
Correct answer	5.3	9.7	10.2	14.1	8.4

Function of the Tests—These tests fill a very definite need for a standardized test in problem solving. The child may be able to perform the separate operations but may not know how to apply them to the solving of problems. Such application of the various operations is the practical measure of school success. A larger element of reasoning is required in applying mathematical opera-

tions to problems than in merely performing the operations themselves; not only is specific ability in arithmetic requisite, but considerable general intelligence is required in solving problems. Although theoretical difficulties may be involved in such combination, they may be neglected in the consideration of the practical values of such tests as these.

The Stone Standardized Reasoning Test in Arithmetic —Test II [2]

This test consists of twelve problems in arithmetic involving thought or reasoning rather than complex mathematical operations. Only the simpler processes of whole numbers and fractions are involved in solving any of the problems but some of the problems are difficult enough to tax the ability of high school pupils. A preliminary exercise is provided for practice, after which fifteen minutes is allowed for the test. The purpose of the test is to measure the ability of pupils to think in terms of mathematical values.

Scoring the Test—The test may be scored in any one of four different ways and norms are provided for each method. If the test is to be used for survey purposes only (1) correct answers should be counted. If it is to be used for diagnostic purposes the solutions may be scored on the basis of (2) reasoning *not counting* partial solutions (3) reasoning *counting* partial solutions, or (4) accuracy of reasoning. In scoring by "correct answers" each solution is checked for correctness of answer and these answers weighted as follows:

[2] Test I is similar to Test II but is an earlier form and not so difficult.

PROBLEM	SCORE	PROBLEM	SCORE
1	1.0	7	1.2
2	1.0	8	1.6
3	1.0	9	2.0
4	1.0	10	2.0
5	1.0	11	2.0
6	1.4	12	2.0

For example, a paper with all twelve answers correct should receive a score of 17.2. For this method of scoring the following norms are given:

STANDARD MEDIAN SCORES ON BASIS OF CORRECT ANSWERS FOR STONE REASONING TEST

School Grade	5	6	7	8
Before diagnostic testing and remedial teaching	4.0	5.0	6.5	7.75
After diagnostic testing and at least three months remedial teaching	5.0	6.0	7.0	8.25

The manual describes in detail how to score the problems by each of the other three methods and gives norms for each.

Function of the Test—This test is a measure of the capacity to solve problems in arithmetic. In so far as this is a reasoning process the test may be called a measure of reasoning ability. From this it is not to be inferred that this is the only kind of reasoning there is. Neither is it probable that there is any general faculty of reasoning as the older psychology would try to make us believe. On the other hand, there may be many elements in common in any process of reasoning. Whether or not this test is a satisfactory measure of any such general capacity remains to be demonstrated. At present we are only justified in saying that it is a fairly successful measure of a pupil's capacity to solve problems in arithmetic not involving difficult numerical computation.

The First Four Problems of the Stone Reasoning Test

Reasoning Test II

Solve as many of the following problems as you have time for; work them in order as numbered.

1. A man starting on a journey took $200. He paid for railroad fare $67; for berth in sleeping car, 4 days, $2 a day; hotel bills, 15 days, $3 a day; other expenses, $25. How much money had he left?
2. Sam had 12 marbles. He found 3 more and then gave 6 to George. How many did Sam have left?
3. A man bought 163 barrels of flour at $11 a barrel. Fifteen barrels were spoiled and the remainder sold at $13 a barrel. Did he gain or lose and how much?
4. A drover bought 132 head of cattle at $45 a head and 67 at $61 a head. He sold them at $50 a head. Did he gain or lose and how much?

The purpose for which the test is to be used should determine which of the four possible methods of scoring should be used. The author suggests that when scored by the "correct answer method" and used together with tests such as the Courtis Research Tests in the fundamental operations, it becomes a valuable measure of arithmetical ability. When scored by the "reasoning not counting partial solution" method, it becomes an important class diagnostic test; when scored by the "reasoning counting partial solution," it becomes a satisfactory measure of the ability of the individual pupil.

Materials Needed

Courtis, S. A., Standard Research Tests in Arithmetic—Series B—for grades 3 to 8. Complete material for testing a class of forty pupils 60 cents, S. A. Courtis, 1806 East Grand Blvd., Detroit, Mich.

Courtis, S. A., Standard Practice Tests in Arithmetic, 48 graded lessons, two forms, A and B. For grades 4 to 8. Cabinet I,

576 lesson cards with guides for a class of 48, price $8.50, Cabinet II, 288 lesson cards with guides for a class of 24, price $6.50, Cabinet III, 144 lesson cards for a class of 12, price $2.25. World Book Co., Yonkers, N. Y.

Monroe, W. S., General Survey Arithmetic Tests, Forms 1 and 2. Scale I is for grades 3, 4, and 5, and scale II for grades 6, 7, and 8. The Public School Publishing Co., Bloomington, Ill. Sample set 15 cents, price $1.00 per hundred.

Monroe, W. S., Standardized Reasoning Tests in Arithmetic, Forms 1 and 2. Test I is for grades 4 and 5, test II for grades 6 and 7 and test III for grade 8. Public School Publishing Co., Bloomington, Ill. Sample set 8 cents, price 80 cents per hundred.

Judd, C. H., The Cleveland Survey Arithmetic Tests for grades 3 to 8. Public School Publishing Co., Bloomington, Ill. Sample set 10 cents, price $1.90 per hundred.

Stone, C. W., Standardized Reasoning Test in Arithmetic, Forms 1 and 2 for grades 5 to 8. Teachers College, Columbia University, Manual 65 cents, price 40 cents per hundred.

Woody, C., and McCall, Wm. A., Mixed Fundamentals, Forms I and II for grades 3 to 8. Teachers College, Columbia University. Sample set 10 cents, price 60 cents per hundred.

Selected References

Freeman, F. N., *The Psychology of the Common Branches,* Chapter IX on Mathematics, Houghton Mifflin Co.

Thorndike, E. L., *The Psychology of Arithmetic,* Teachers College, Columbia University.

Woody, C., *The Measurement of Some Achievements in Arithmetic,* Teachers College, Columbia University.

CHAPTER X

GEOGRAPHY

The construction of a standard test in a content subject like geography is much more difficult than the construction of a test in a formal subject like spelling. One of the chief reasons is the difficulty of determining the essentials to be taught in a subject like geography. A group of geography teachers no doubt would not agree upon the relative amount of time to be spent on memory work and thought problems, especially in the lower grades; or on the other problems such as the relative importance of home geography, earth history, form changes, knowledge and use of maps, ability to deduce and use geographical principles and other similar problems. These are all problems that the teacher will admit are important, but lack of time prevents an adequate consideration of all of them. Which will the teacher emphasize and which should she neglect?

In the case of reading or spelling there is an answer to such problems as has been pointed out in the chapters on these subjects. In the content subjects these problems still remain unsolved and so they are often made the subject of extended and generally unprofitable debate.

There have been several attempts at the construction of tests and scales in geography. A few of the more successful of these will be described.

The Hahn-Lackey Geography Scale

Description of the Scale—This is one of the oldest and in some respects the most satisfactory geography scale

V	W	X	Y	
79	84	88	92	FOURTH GRADE
88	92	94	96	FIFTH GRADE
92	94	96	98	SIXTH GRADE
96	98	99	100	SEVENTH GRADE
96	98	99	100	EIGHTH GRADE

63. What direction do you live from the equator?

64. Name two important mountain ranges of the United States

11 Name the four seasons.

33. What is the direction half way between south and west?

3. In what direction would you go to go to Canada?

15. What is the capital of the United States?

25. Name an animal useful to man in desert countries.

65. Why is there not much farming done in Alaska?

23. How do we know that there is air?

4. Why is the Arctic Ocean not used much by sailors?

40. How does the ocean help to furnish us food?

36. Why are there more birds here in summer than in winter?

23 In what direction are you facing when your back is toward the north?

51. What is the capital of your state?

17. Name two things plants must have to live.

2. Name two animals used by the Eskimos.

Hahn-Lackey Geography Scale

DIRECTIONS FOR GIVING TESTS

1. Do not impose a time limit.

2. Select from four to ten exercises from the list given in any one column. Write your selection of exercises for the test on the blackboard and proceed as in an ordinary school examination.

3. Give the following instructions to the pupils taking the test: "You will be given enough time to answer as many of these exercises as you can. Kindly read each exercise carefully to get its exact meaning, then write the best answer you can in the fewest words. Complete sentences are not necessary, words or phrases will do. You are not expected to be able to answer all the exercises Some of them were made difficult on purpose, but if you can answer the difficult ones, the credit due you will be that much greater. At any rate, please try hard to answer every exercise. Ask no questions about any of the exercises in the test. If your teachers should permit you to ask questions and then answer them for you, it would defeat the purpose of the test and your answers could not be used."

4. In scoring answers to exercises consisting of two or more parts, give credit for each part answered correctly. In general, give credit for an answer that clearly indicates a knowledge of the idea involved in the exercise. Copies of directions indicating specifically what to accept and what to reject in scoring answers may be obtained from the authors for Five Cents apiece.

5. The standards given are correct for 1922.

FIG. 12. A SECTION OF THE EASIER END OF THE HAHN-LACKEY GEOGRAPHY SCALE

yet devised. It consists of a series of 217 questions arranged on the same general plan as the Ayres Spelling Scale, that is, the questions are arranged in twenty-three

columns with from one to eighteen questions in each column. The questions in any column are of approximately equal difficulty and the columns are arranged in order of difficulty. (It differs from the spelling scale in that the easier questions are at the right or "X" end of the scale.) Norms for the different school grades are given at the top of each column. For example, a fifth grade pupil should make a score of 66 in column "R" or a score of 73 in column "S" of the scale.

The questions in this scale are about equally divided between "memory ability" and "thinking ability." The questions of the former type are printed in light face and the latter in black face type. In order to make the scale diagnostic the author has classified the questions under seven other headings. These are:

1. Knowledge of home geography.
2. Knowledge of the meaning of technical terms or symbols.
3. Knowledge of the map as a geographical tool.
4. Ability to locate places in geography.
5. Ability to use constructive imagination to see geographical situations as they are.
6. Ability to think inductively or derive general principles.
7. Ability to think deductively or deduce geographical from general principles.

Method of Using the Scale—The teacher may select a list of questions from one of the columns for use with the class. The authors suggest that from four to ten questions should be selected. These questions may be

meaning or thought questions as the teacher chooses. The pupils are allowed as much time as they need for answering these questions. The answers are graded by means of a key which is furnished with the scale on the basis of 100 as a perfect score. The grade for a pupil or the average grade for a class may be compared with what is to be expected by reference to the norms given at the top of the column from which the questions were selected.

Function of the Scale—Such a list of questions as this provides the teacher with a profitable scource of material for testing her pupils. But as a scale it has certain features which limit its usefulness. The scale must be used wisely or there are many possibilities that the questions will be selected improperly. The teacher may select all memory questions and justify her teaching by the high scores of her pupils on them. The method of selecting the questions permits the teacher to select the ones her pupils can answer and omit the ones they can not answer. On the other hand she may so select the questions that they are unfair for her class. For these reasons too much importance should not be placed on the norms as given. They are valuable indications of what should be expected and if the questions are well selected and carefully graded they are reliable measures of geographical knowledge. The diagnostic feature of the scale also adds to its usefulness.

The Posey-Van Wagenen Geography Scales

Description of the Scales—The Posey-Van Wagenen Scales, of which there are ten in all, each consist of a series of thirty questions. In some of the scales the

questions are thought questions and in others they ar information questions. "The material for these geo raphy scales has been selected as representative of wide range of interests." "The selection of the mater the framing of the tasks, and their final arrangement int scales is the result of judgments based upon a wide rang of geographical information and a wide experience i scale making and using." Four of the tests are for grade 5 and 6 and the others are for grades 7 and 8.

Method of Using the Scales—The teacher selects the scale which tests the grade and subject matter that it i desired to test. The pupils are given forty minutes t answer as many of the thirty questions of the scale a they can. This is adequate time for most of them to answer as many questions as they know the answers for. Keys are given in the manual for scoring the different scales. The scales are each separated into three parts for scoring, each part consisting of ten questions. The scores for each of these three parts are kept separately. The pupil's final score is found by the use of a somewhat complicated set of tables to be found in the manual. Norms for the different school grades are given for the end of the year. These norms apply for any one of the scales.

THE POSEY-VAN WAGENEN GEOGRAPHY NORMS

School Grade	5	6	7	8
Norm	68	74	80	86

Function of the Scales—No one of these scales cover a wide range of geographical knowledge but each calls f fundamental and representative m al in geograph This is shown by the fact that the probable error is on

Posey-Van Wagenen Geography Scales

Information R (General), Division 1. *Grades* 5 *and* 6

GROUP II (*Average value 69.5*)

(65)
On what do each of these things grow: a vine, a bush, a plant, a tree?
1. Olives .....
2. Dates .....
3. Figs .....
4. Rice .....

16.(70)
(a) Is winter wheat sown in the spring, summer, fall or winter?
.....
(b) Is it harvested in the spring, summer, fall, or winter?
.....

Posey-Van Wagenen Geography Scales

Thought S, Division 1. *Grades* 5 *and* 6

5.(59)
Why does not the palm tree grow in Canada?
.....

6.(60)
It costs very much less for the same weight of any material to be carried by boat than by train. At the same time the train takes less time to go between the same two places.
(a) If you had plenty of time and very little money, and had to go from New York to New Orleans, how would you go?
.....

(63)
In which of these regions would you expect to find the largest number of people living: a fertile plain, a mountainous territory, an indented coast line open to the interior, a desert, an iron mining region?
.....

10.(64)
It is much easier and cheaper to restore fertility to the soil by letting the land lie fallow or unused for a few years than to buy fertilizer or manure to put on it.
Is the land in sparsely settled regions more likely to be fertilized or allowed to lie unu[illegible]
.....

2.1 for a score on these scales. This means that in half the scores the pupil's actual ability in that phase of the subject tested does not vary from the scale scores by more than 2.1 points or units on the scale. These units (called deci-quartiles) are approximately one-tenth of the amount of gain which should be expected of normal pupils during a year of the elementary school. With the use of this unit the scales are so constructed that a difference of 5 points between any two places on the scale is the same as five points difference in any other part of the scale. Not only is this true, but a score of 60, for example, represents the same value on any one of the scales for the fifth and sixth grades and may be compared with the score of a pupil on any of the seventh and eighth grade scales. In other words, the units are uniform throughout all the scales and directly transferable from one grade to another. This is a unique feature in scale construction.

The scales are easy to give but the scoring is somewhat complicated. This feature of the scoring is necessary to make the comparisons between the different scales possible. While the separation of the scales into thought scales and information scales prolongs the testing program, it makes them of greater diagnostic value. In general these are the most satisfactory geography scales yet devised.

The Buckingham-Stevenson Place Geography Tests

Description of the Tests—The tests consist of two parts, one a series of tests on knowledge of the location of places in the world and the other of location of places

in the United States. In the first test there are 70 locations to be made. Some of these are countries, mountains, lakes or rivers to be located by naming the continent "on or nearest to which each" is located. Another portion of the test consists of a list of seaports for which the name of the nearest ocean is to be given. Ten cities are to be located by naming the country in which each is located.

In the place geography tests for the United States fifteen cities are to be located by states, a list of states is named and the pupil is asked to name the state north, west, east or south of the given state. There are sixty questions in this test. Three forms of each of the two tests are available.

Method of Using the Tests—The questions are to be read by the teacher and the answers written by the pupil. Sufficient time is allowed for all pupils to write the answers. Before the test is given a preliminary trial exercise is studied to familiarize the pupil with the nature of the test. Preliminary standard scores are given for each form of the two tests as follows:

THE BUCKINGHAM-STEVENSON PLACE GEOGRAPHY TESTS

The World

Grades	8		7		6		5		4
	high	low	high	low	high	low	high	low	high
Form 1 Score.....	44.0	54.0	50.3	48.0	39.0	33.4	39.0	26.3	14.1
Form 2 "	47.7	52.6	53.7	35.5	29.5	32.9	36.6	29.7	16.0
Form 3 "	46.9	51.7	52.9	39.2	34.5	37.8	39.0	37.5	16.9

The United States

Form 1 Score.....	31.6	35.0	34.9	28.3	36.2	13.5	17.6	12.8	16.7
Form 2 "	29.5	34.5	32.0	27.0	36.5	27.0	14.8	14.7	16.3
Form 3 "	32.7	41.0	37.4	30.2	38.5	29.0	17.3	15.1	14.6

Function of the Tests—These tests are exceptionally satisfactory for the purpose for which they are intended. They measure the pupils' information about locations in geography. They do no more than this and are not intended to measure other types of geographical information. The tests are simple and easy to give and score. Any teacher can give the tests without special preparation or material other than a sample of the tests. They do not take up a large amount of time and still they cover sufficient subject matter to represent a fair sample of the pupil's ability in place geography. It is to be hoped that similar tests will be devised for other types of geographical information.

The Courtis Location Geography Tests

Description of the Tests—These tests are somewhat similar to the Buckingham-Stevenson Place Geography Tests. There are two tests, one for locations of oceans, continents and countries of the world; the other for the location of states and important cities of the United States. In each test a map is shown with the different parts indicated by numbers. For example, in the map of the United States each state has a number. Test B consists of a list of 43 countries of the world to be located by continents. The pupil is to place the number of the continent in which the country is found after its name. The same map is to be used for the location of seven continents and five oceans. The test for the United States is very similar except that states and important cities are to be located.

Method of Using the Tests—The pupils are furnished copies of the test and after a preliminary exercise they

are given one minute for finding as many answers as possible to the continents and oceans problems and four minutes for the location of countries. This is for the World Test. For the United States tests four minutes is allowed for the states and two minutes for cities. The score is a weighted per cent with 1,000 as the highest possible score in each part of each test.

Function of the Tests—These tests are simple, brief and easy to give. The use of the maps probably adds to the value of the tests but they are not as comprehensive as the Buckingham-Stevenson tests. The lack of norms seriously limits the usefulness of the tests for the classroom teacher.

Materials Needed

Buckingham, B. R., Stevenson, P. R., Place Geography tests. One tests "The World" and the other "United States." Three forms of each for grades 4 to 8. Only one copy of the test needed as the teacher dictates the test. Public School Publishing Co. Booklet containing 3 forms each of both tests, 15 cents. Class record sheets, each 1 cent.

Curtis, S. H., Standard Supervisory tests—Geography—Location for grades 4 to 8. S. A. Courtis, 1807 East Grand Blvd., Detroit, Mich. Price $1.50 for materials for a class of 40 children.

Hahn, H. H.; Lackey, E. H., Geography Scale for grades 4 to 8. Only one copy needed. Public School Publishing Co., Bloomington, Ill., price 20 cents.

Posey, C. J.; Van Wagenen, M. J., Geography Scales: Teachers Handbook 20 cents. Sample set, 30 cents, $1.50 per 100. Scale Thought S Div. I (grades 5 and 6). Scale Thought R Div. II (grades 7 and 8). Scale Information R (gen.) Div. I (grades 5 and 6). Scale Information R (gen.) Div. II (grades 7 and 8). Scale Information S (gen.) Div. I (grades 5 and 6). Scale Information S (gen.) Div. II (grades 7 and 8). Scale Information A (U. S. and North America) Div. I (grades 5 and 6). Scale Information A (U. S. and North America) Div. II (grades 7 and 8). Scale Information F (Europe) Div. II (grades 7 and 8). Scale Information K (S. A., Asia, Africa) Div. II (grades 7 and 8). Public School Publishing Co., Bloomington, Ill.

Selected References

Courtis, S. A., "Measuring the Effects of Supervision in Geography," *School & Society*, July 19, 1919.

Freeman, F. N., *The Psychology of the Common Branches*, Geography, Chapter VIII, Houghton Mifflin Co.

Lackey, E. E., "A Scale for Measuring Ability of Children in Geography," *Journal of Educational Psychology*, October, 1918.

Posey, C. J. and Van Wagenen, M. J., "The Posey-Van Wagenen Geography Scales," *Teachers Handbook*, Public School Publishing Co., Bloomington, Ill.

CHAPTER XI

HISTORY

In the construction of tests and scales, history presents the same type of difficulty as geography. There is no general agreement as to what should constitute the content of the course and as yet no method has been devised by which the problem may be solved objectively. The older methods of teaching history placed great emphasis upon facts as such, especially upon wars and the lives of kings. The present tendency, in so far as there is a tendency, is to place greater emphasis on causes, historical sequence, economic, industrial and political movements with an interpretation of the present in terms of the past.

In some of the tests these different problems have been given separate consideration. In others questions have been grouped together without attention to the type of information called for. Five of the more widely known and useful tests will be described.

The Hahn History Scale

Description of the Scale—This scale is constructed on the same general principles as the Hahn-Lackey Geography Scale. It consists of about 275 questions in history arranged in columns. The questions in any column are of approximately equal difficulty. Standard scores for the eighth grade are given at the top of each column.

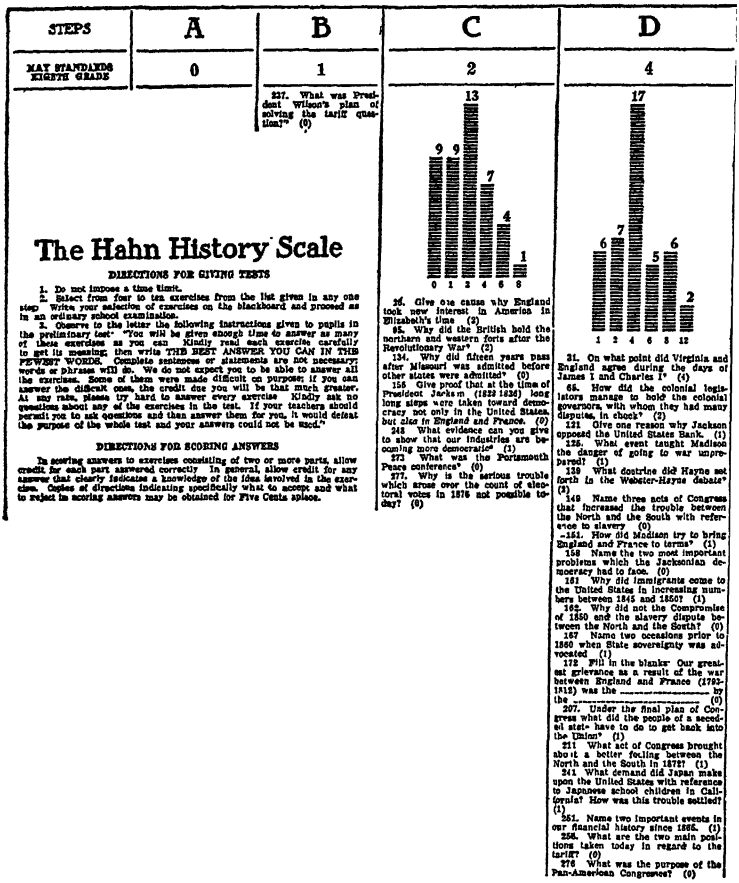


FIG. 13. A SECTION OF THE HAHN HISTORY SCALE

The scale is also for use with the seventh grade. Following each question is a figure giving its value in a seventh grade test. The overlapping in the grades, that is, the per cent of a class that is only equal to the next lower or is equal to the next higher grade in school, is

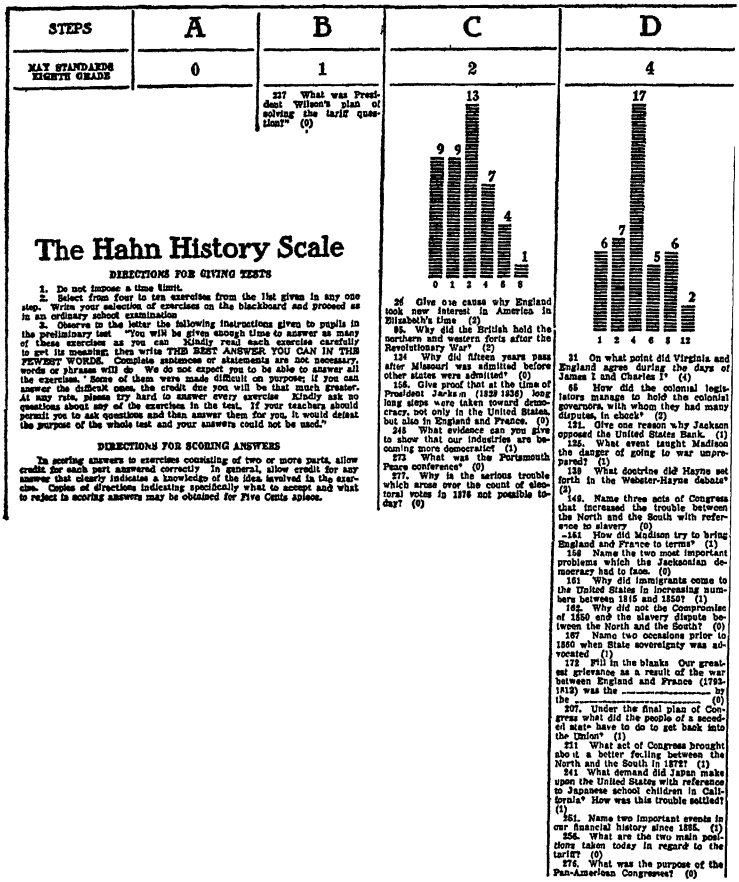
indicated by graphs at the top of each column. These graphs were constructed from results obtained from 43 schools in 14 states. For example, for the eighth grade 14 classes out of the 43 made the normal grade of 21 in column I, 8 classes made a score of 16, the normal score for column "H," and 5 classes made a score of 12, the normal score for column "G." Ten other classes made a score 27, the normal score for column "J," and 6 classes made a score of 34, the normal score for column "K."

The scale is diagnostic in that the questions have been classified under nine different abilities. These range from memory ability for historical facts to ability to see connections and make historical comparisons and judgments.

Method of Using and Scoring—The teacher in the eighth grade should "select from four to ten exercises from any one step" on the scale. These exercises may be written on the blackboard and the pupils given all the time they need in answering the questions. The papers are to be graded by the usual method of scoring on the basis of 100 points for a perfect score. Correct answers to each question are given in a key which is furnished with the scale.

The seventh grade teacher may select either a list of questions with the same seventh grade values or a list with different values and average these values for the seventh grade standard. For example, if four questions have seventh grade values—as given after each question —of 38, 40, 40 and 46, a seventh grade class should make a score of 41 on this test.

Function of the Scale—In general the same criticism



STEPS	A	B	C	D
MAY STANDARDS EIGHTH GRADE	0	1	2	4

227 What was President Wilson's plan of solving the tariff question?" (0)

The Hahn History Scale

DIRECTIONS FOR GIVING TESTS

1. Do not impose a time limit.
2. Select from four to ten exercises from the list given in any one step. Write your selection of exercises on the blackboard and proceed as in an ordinary school examination
3. Observe to the letter the following instructions given to pupils in the preliminary test "You will be given enough time to answer as many of these exercises as you can Kindly read each exercise carefully to get its meaning; then write THE BEST ANSWER YOU CAN IN THE FEWEST WORDS. Complete sentences or statements are not necessary; words or phrases will do We do not expect you to be able to answer all the exercises. ' Some of them were made difficult on purpose; if you can answer the difficult ones, the credit due you will be that much greater. At any rate, please try hard to answer every exercise Kindly ask no questions about any of the exercises in the test. If your teachers should permit you to ask questions and then answer them for you, it would defeat the purpose of the whole test and your answers could not be used."

DIRECTIONS FOR SCORING ANSWERS

In scoring answers to exercises consisting of two or more parts, allow credit for each part answered correctly In general, allow credit for any answer that clearly indicates a knowledge of the idea involved in the exercise. Copies of directions indicating specifically what to accept and what to reject in scoring answers may be obtained for Five Cents apiece.

26 Give one cause why England took new interest in America in Elizabeth's time (2)

95. Why did the British hold the northern and western forts after the Revolutionary War? (2)

124 Why did fifteen years pass after Missouri was admitted before other states were admitted? (0)

156. Give proof that at the time of President Jackson (1829 1836) long long steps were taken toward democracy, not only in the United States, but also in England and France. (0)

245 What evidence can you give to show that our industries are becoming more democratic? (1)

273 What was the Portsmouth Peace conference? (0)

277. Why is the serious trouble which arose over the count of electoral votes in 1876 not possible today? (0)

21 On what point did Virginia and England agree during the days of James I and Charles I? (4)

65 How did the colonial legislators manage to hold the colonial governors, with whom they had many disputes, in check? (2)

121. Give one reason why Jackson opposed the United States Bank. (1)

125. What event taught Madison the danger of going to war unprepared? (1)

139 What doctrine did Hayne set forth in the Webster-Hayne debate? (3)

142. Name three acts of Congress that increased the trouble between the North and the South with reference to slavery (0)

-151. How did Madison try to bring England and France to terms? (1)

158 Name the two most important problems which the Jacksonian democracy had to face. (0)

161 Why did immigrants come to the United States in increasing numbers between 1815 and 1850? (1)

162. Why did not the Compromise of 1850 end the slavery dispute between the North and the South? (0)

167 Name two occasions prior to 1850 when State sovereignty was advocated (1)

172 Fill in the blanks Our greatest grievance as a result of the war between England and France (1793-1813) was the _____ by the _____ (0)

207. Under the final plan of Congress what did the people of a seceded state have to do to get back into the Union? (1)

211 What act of Congress brought about a better feeling between the North and the South in 1872? (1)

241 What demand did Japan make upon the United States with reference to Japanese school children in California? How was this trouble settled? (1)

251. Name two important events in our financial history since 1895. (1)

256. What are the two main positions taken today in regard to the tariff? (0)

276. What was the purpose of the Pan-American Congresses? (0)

FIG. 13. A SECTION OF THE HAHN HISTORY SCALE

The scale is also for use with the seventh grade. Following each question is a figure giving its value in a seventh grade test. The overlapping in the grades, that is, the per cent of a class that is only equal to the next lower or is equal to the next higher grade in school, is

indicated by graphs at the top of each column. These graphs were constructed from results obtained from 43 schools in 14 states. For example, for the eighth grade 14 classes out of the 43 made the normal grade of 21 in column I, 8 classes made a score of 16, the normal score for column "H," and 5 classes made a score of 12, the normal score for column "G." Ten other classes made a score 27, the normal score for column "J," and 6 classes made a score of 34, the normal score for column "K."

The scale is diagnostic in that the questions have been classified under nine different abilities. These range from memory ability for historical facts to ability to see connections and make historical comparisons and judgments.

Method of Using and Scoring—The teacher in the eighth grade should "select from four to ten exercises from any one step" on the scale. These exercises may be written on the blackboard and the pupils given all the time they need in answering the questions. The papers are to be graded by the usual method of scoring on the basis of 100 points for a perfect score. Correct answers to each question are given in a key which is furnished with the scale.

The seventh grade teacher may select either a list of questions with the same seventh grade values or a list with different values and average these values for the seventh grade standard. For example, if four questions have seventh grade values—as given after each question —of 38, 40, 40 and 46, a seventh grade class should make a score of 41 on this test.

Function of the Scale—In general the same criticism

may be made of this scale as the Hahn-Lackey Geography Scale. It provides the history teacher with a large list—probably not a "complete" list as the author suggests—of fairly well standardized questions for use in seventh and eighth grades in history. The material from which the questions were taken is common to six modern texts in history. The diagnostic feature of the test gives the teacher an opportunity to determine in which types of ability the pupils are well prepared and the types in which they are deficient. No attempt has been made to indicate the relative importance of these different types of abilities. While the scale is not too difficult for use by the classroom teacher if she will follow the directions carefully, it seems likely that the same material would be more usable if made up into a series of alternative tests with definite norms.

The Barr Diagnostic Tests in American History

Description of the Tests—Recognizing the fact that the content of history is not standardized, Barr has constructed a test for each of five "fundamental processes" in history. These five processes are:

1. Comprehension of historical facts.
2. Chronological judgment.
3. Historical evidence.
4. Time relation of events.
5. Causal relationship in history.

Following a practice test are the five series of tests. The first consists of paragraphs chosen from various sources of history followed by sets of questions. In the second, lists of persons or events are to be arranged in chronological order. In the third, various sources of

history are to be weighed as to their importance or reliability. In the fourth, series of historical events are to be arranged in order of importance. In the fifth, historical events are to be connected with their causes.

Six minutes is allowed for each test. There are two forms of the tests, 2A and 2B, of which 2A is the more difficult. An answer sheet is provided as an aid in scoring the papers.

Scoring the Tests—Each question answered has a certain weight or score value. This value is given after the answer to the question on the scoring sheet. A pupil's score is the sum of the weights of all the questions answered correctly. These scores are kept separately for each of the five tests. Tentative norms are given as follows:

Test	I	II	III	IV	V
Scale 2B—8th grade	7.2	7.1	5.7	9.6	5.4
Scale 2B—Senior high school	10.2	8.0	9.0	12.3	7.5
Scale 2A—Senior high school	8.2	7.5	6.2	12.2	8.9

Function of the Tests—These tests present a very interesting attempt to solve the problem of tests in history. The separation of the material into five different tests gives them an important diagnostic value. It also makes possible a differentiation of the methods of teaching to meet individual differences. The material of the tests does not cover in any sense the whole field of American history but that which was chosen is important and representative. Some authorities might differ from the author in the answers to some of the questions but in general the answers are unequivocal. The most serious criticism is the complexity in the type of ability called for by the tests. This ranges from an almost purely

Sample Page from the Barr Diagnostic History Tests

TEST III

7. Following is a letter from Governor Gibson:

Vincennes, Indiana Territory,
August 10, 1812.

"Col. Wm. Hargrave,
Commanding Mounted Rangers:
"Two scouts from this post were at a point on the west White river thirty miles east of the forks and saw two old Delaware Indian men who have a lone wigwam at that place. These Indians were friendly and have been for a long time. They said that several Pottawattamies had recently been at that point and told them—'soon we will go to the Ohio river—get heap horses—maybe get scalps—the British drive Americans away soon.'

John Gibson,
Acting Governor."

Put a cross (X) before each of the following questions that you would like to have answered if you were writing a history of the War of 1812 and came upon the above account.

(a) What was the training and profession of the writer?
(b) Was the writer prejudiced?
(c) Was the author's literary style good?
(d) Was the writer in a position to know the facts?
(e) Did the writer take the trouble to get the facts?

End of Test III. Check your answers and then wait quietly until all finish.

TEST IV

1. Put a cross (X) before the event in the following list which has been of the greatest importance in American History.
   (a) Braddock's defeat
   (b) Burr's conspiracy
   (c) The Hayes-Tilden contest
   (d) The discovery of America
   (e) The Webster-Hayne debate

2. Put a cross (X) before the event in the following list which has been of the greatest importance in the economic development of the United States.
   (a) The Tariff Act of 1832
   (b) The invention of the telephone
   (c) The panic of 1873
   (d) The laying of the Atlantic cable
   (e) The introduction of railway transportation

reading test, in which the subject matter is historical information, to tests in judgment, depending more on general intelligence than upon historical information. This has certain advantages as has already been pointed out but violates one general principle of tests, that of unity and definiteness of the subject matter. The time limit of six minutes for each test introduces the factor of speed of reading and speed of writing as factors that may be measured rather than, or at least along with, the measure of ability in history. The derivation of more reliable norms would add to the value of the tests.

The Harlan Test of Information in American History

Description of the Test—This test consists of ten exercises for measuring different types of historical knowledge. Each exercise, except No. IV, contains from four to six questions dealing with men, events and dates in American history. Exercise IV contains two questions in civics. The pupil is allowed as much time as necessary to finish the test. Most pupils finish in twenty-five minutes.

Scoring the Test—The scoring is done with the aid of a scoring key. Each element of each exercise receives a score of 2, 1, or 0, according to whether it is correct, half correct, or entirely wrong. There are fifty elements in the test and 100 constitutes a perfect score. Median scores for the end of the school year are given as follows:

HARLAN AMERICAN HISTORY TEST

School Grade	7	8
Norm	56	86

Function of the Test—This test is simpler than the two scales previously described and, therefore, it is prob-

The Last Page of the Harlan Test of Information in American History

EXERCISE VIII. Score.....

Below are some general statements concerning the history of our country. Prove that they are true by stating a typical example or instance in American History which has shown them to be true.

1. *One method employed by a nation in acquiring territory is by conquest.*
   .....
2. *The final decision of civilized people is that the enslavement of one people by another is wrong.*
   .....
3. *The national congress has regarded unrestricted immigration as dangerous to the welfare of the nation.*
   .....
4. *An exaggerated idea of the power of the president has, at times, endangered the life of the president.*
   .....

EXERCISE IX. Score.....

The following topics represent matters of importance in the history of the United States. State definitely of what significance each has been.

1. *Articles of Confederation*.....
   .....
2. *Mason and Dixon's line*.....
   .....
3. *Monroe Doctrine*.....
   .....
4. *The Tariff*.....
   .....

ably better suited for ordinary classroom use. The teacher is not required to select questions for use and yet a wide range of historical knowledge is tested. The subject matter is well standardized as shown by the fact

that it was based upon the Bagley and Rugg [1] study of the content of twenty-three standard textbooks in American history. The test would be more valuable if there were several alternative forms.

It is diagnostic since each exercise tests a different type of historical ability. The seventh or eighth grade teacher of history can well make use of this test in order to determine the standing of her pupils as compared with what normal pupils should know in American history.

The Van Wagenen American History Scales—Revised Edition

Description of the Scales—These scales are constructed on the same general principle as the Posey-Van Wagenen Geography Scales. There is one scale for information and one for thought. The information scale consists of a series of questions in American History some of which are to be answered by checking the right answer from a list of possible responses, others by writing the correct answer. There are thirty questions and they cover a wide range of historical information. The thought scales consist of lists of historical facts or events followed by a question regarding the cause or reasons for the fact or event. The answer to the question is not given in the material read but is to be deduced from the material given plus the pupil's wider experience with historical facts and life. The answers are a matter of judgment on the part of the pupil. There are also thirty questions in this scale. Division 1 of each scale is for grades 5 and 6 and division 2 for grades 7 and 8.

[1] W. C. Bagley and H. O. Rugg, *The Content of American History as Taught in the Seventh and Eighth Grades*, U. of Ill. School of Education Bulletin, No. 16, 1916.

Questions Selected from the Van Wagenen American History Scales—Revised Edition

From Information. *R. General Division 1. Grades 5 and 6. Group I*

3. (56) Put a check mark in front of each of these things which were in use during the Civil War.

..... Submarine
..... Poison Gas
..... Cavalry
..... Ironclad war vessels
..... Aëroplanes

4. (57) What were the first four European countries to make settlements in America?

1 .....
2 .....
3 .....
4 .....

From Thought. *R. For Grades 7 and 8. Group II*

12. (77) Previous to the Civil War a large part of the Southern cotton crop was exported to England.

What was evidently one of the chief occupations of England?

13. (78) In 1800, Spain gave Louisiana up to France. The United States, fearing that France might set up a colony and control the Mississippi River, was anxious to get Louisiana. In 1803, Napoleon of France feared that Great Britain was about to seize his American territory.

What would you expect Napoleon to do?

14. (79) In 1810, nine tenths of our foreign trade (980,000 tons) was carried in American vessels. The War of 1812-14 stopped the importation of foreign-made goods.

In what industry would you expect American capital soon to have become invested?

Method of Using the Scales—The pupils are furnished with copies of the scale selected for use. Either the thought scale or the information scale may be given or both may be given. Forty minutes are allowed for the test. A key is provided for use in grading the pupil's answers. The method of scoring is the same as that used in scoring the Posey-Van Wagenen Geography Scales. Each of the three parts of the scale is scored separately. The final score for a pupil is a corrected value obtained from these three separate scores.

Function of the Scales—These scales have separated the thought and information factors in history and as a result the scales become diagnostic. The separation of these two factors seems essential in the construction of a history scale. The subject matter of the scales covers a wide range of historical knowledge and, therefore, the scales give a fair sampling of the pupil's ability. The questions are upon important phases of history and are clear-cut. Very little writing is demanded of the pupil. Most of the answers consist in checking the right answers from a list of possible answers. Sometimes brief statements are to be written by the pupil. The method of scoring the scales, while justifiable from the point of view of statistical method, is somewhat complicated for the average classroom teacher. This seems to be the most serious criticism of the scales.

The Pressey-Richards Tests in the Understanding of American History

Description of the Tests—There are four tests in this group. Test I is a measure of Character Judgment, test II of Historical Vocabulary, test III of Sequence of

Sample Exercises from the Pressey-Richards Tests for the Understanding of American History

Test I—Character Judgment

Complete directions instruct the pupils to underline the one word after each man's name which he thinks best describes him.

14. Aaron Burr: conscientious, honored, shy, disloyal.
15. Daniel Webster: Eloquent, quarrelsome, clever, dominating.
26. Herbert Hoover: Efficient, assertive, nervous, talkative.

Test II—Historical Vocabulary

The pupils are directed to underline the one of the four statements after each question which is the correct answer.

4. What is a confederacy? A disunion, A colony, A commonwealth, A league of states.
5. What is an autocracy? Representative government, Mob law, Self-government, An absolute form of government.
22. What is a panic? A mass of people, A political disturbance, A financial crisis, A gold rush.

Test III—Sequence of Events

Of the four events in each question the pupils are directed to underline the event that happened the longest time ago.

5. First Continental Congress, Hartford Convention, Constitutional Convention, Declaration of Independence.
6. Battle of Yorktown, of Saratoga, of Bunker Hill, of Lexington.

Test IV—Cause and Effect Relationship

Of the four events given in each question three are causes and one the effect. The pupils are directed to underline the effect.

10. Ratification of the Treaty of 1819, Holy Alliance, Need for independence in South America, Monroe Doctrine.
11. Issuing of paper money by state banks, Closing of U. S. Bank, Specie Circular, Panic of 1837.

Events and test IV of Cause and Effect Relationship. Each test consists of twenty-five exercises and one practice exercise.

The Character Judgment test contains the names of prominent persons or events in American history. Following each name there are four adjectives. The pupil is told to underline the adjective after each name which best describes the person or event. The Historical Vocabulary test contains twenty-six historical terms. Each is followed by four possible answers. The pupil is directed to underline the correct answer to each question. The Sequence of Events test consists of twenty-six sets of four historical events. The pupil is directed to underline the event in each set that happened the longest time ago. The Cause and Effect Relationship test consists of twenty-six lists of events, with four in each list, three of which acted as "causes" and one of which was the "effect." The pupil is asked to underline the effect in each list.

Five minutes is allowed for the first test, six minutes for the second, six for the third and eight for the last. One point credit is allowed for each of the twenty-five exercises in each test not counting the practice exercise. The total possible score is, therefore, 100 points. Norms for the test are given as follows:

NORMS FOR THE PRESSEY-RICHARDS TESTS FOR THE UNDERSTANDING OF AMERICAN HISTORY

Grades	6	7	8	12
Total Score	21	29	41	63
Test I	6	7	10	15
Test II	5	7	11	17
Test III	5	8	11	16
Test IV	5	7	9	15

Function of the Tests—These tests are simple in construction, easy to administer and diagnostic in character. This diagnosis goes further than most tests in that it indicates a very successful attempt to evaluate the aims and purposes in the teaching of history. The first test, Character Judgment, gives a measure of the principal aim, especially in the lower grades; the second test measures general historical information, while the third and fourth tests measure those reasoning factors most stressed in the upper grades. Only about thirty minutes is required for the test and this time is ample enough to make it a power test rather than a speed test. From the results of this test the teacher may discover the standing of her pupils and an indication of the sort of things they know as well as those they do not know. The material of the test represents a well-balanced selection from the different fields of history. In brief these tests are a very satisfactory measure of historical information for use by the classroom teacher in American history.

Materials Needed

Barr, A. S., Diagnostic tests in American History, Series B for 8th grade and Senior high school. Public School Publishing Co., Bloomington, Ill. Sample set 15 cents. Price $4.00 per hundred.

Hahn, H. H., The Hahn History Scale for grades 7 and 8. Public School Publishing Co., Bloomington, Ill. One copy sufficient. Price 20 cents.

Harlan, C. L., Test for Information in American History for grades 7 and 8. Public School Publishing Co., Bloomington, Ill. Sample set 6 cents, price 80 cents per hundred.

Pressey, L. W., and Richards, R. C., A Test for the Understanding of American History, for grades 6, 7 and 8, and Senior High

School. Public School Publishing Co., Bloomington, Ill. Sample set 10 cents, $2.00 per hundred.

Van Wagenen, M. J., Reading Scales, History, for grades 5 to 8. Public School Publishing Co., Bloomington, Ill. Scale R Information (General) Division 1 for grades 5 and 6. Scale R Division 2 for grades 7 and 8. Scale Thought R Division 1 for grades 5 and 6, Division 2 for grades 7 and 8. Prof. M. J Van Wagenen, University of Minnesota, Minneapolis, Minn.

Selected References

Bagley, W. C. and Rugg, H. O., *The Content of American History as Taught in the Seventh and Eighth Grades,* University of Ill., School of Education, Bulletin No. 16, 1916.

Tryon, R. M., *Teaching History in the Junior and Senior High School,* Scott Foresman Co., New York City.

Twenty-second Yearbook of the National Society for the Study of Education, Part II, Public School Publishing Co., Bloomington, Ill.

CHAPTER XII

MUSIC

Music differs greatly from most of the other school subjects in that it is a rather specialized ability. Among most of the school subjects there are many elements in common. Despite the popular opinion to the contrary Starch has shown that there is a high correlation between success in the different school subjects. That is, if a boy makes high grades in arithmetic he is more likely than not to make high grades in history and in language. In general the same holds true for music. On the other hand, success in music sometimes may have little relation to success in other school subjects. The writer well remembers a recital given a few years ago by the inmates of one of the institutions for the feeble-minded, and the performance was much better than the average for public orchestral performances. A girl unable to do her school work because of limited scholarship ability is composing songs for her school of four hundred pupils located in the better section of a medium sized city. It is likely that the average intelligence of choruses of even the best of our musical comedy troupes is not high. Yet it is not intended to intimate that our composers and musicians have been or are inferior men. As a class they are surely much above the average. A Bach, a Schubert, or a Wagner are examples of the case in point. Paderewski has shown himself to be a statesman as well as a musician.

A girl prodigy in music, aged eight, was recently given an intelligence test and made a score equal to a child of eleven years. These are only a few concrete illustrations of the fact that there may or may not be a close correlation between musical and school abilities.

Musical ability itself is no doubt a complex of many specific abilities such as tone production, pitch discrimination, sense of time, rhythm and harmony as well as acquaintance with the symbolism used in music. These more specific abilities do not enter into all types of music with the same degree of importance. For example, the singer is more concerned with tone production, quality, time and rhythm, while the pianist may pay less attention to some of these and be more concerned with the mechanics of tone production.

Music, as one of the so-called fine arts, is looked upon by many as dependent upon an ability present in some persons and absent in others. In a sense this is true just as it is with intelligence. But like intelligence, musical ability exists to some degree in most of us. Our knowledge of music depends upon the amount of training that has been given this native capacity. In most of our better school systems musical training is given. Just as it is important to know the general intelligence of a child in order most wisely to direct his general education, so likewise is it worth while to know the native ability of the child in the field of music. Fortunately we are provided with a very reliable measure of this ability.

The Seashore Measure of Musical Talent

Description of the Tests—These tests of native musical ability consist of a series of five phonograph records

especially constructed for this purpose. The first record is a test in Pitch Discrimination. One hundred pairs of tones are sounded by playing this record on an ordinary phonograph. The pupil or class is furnished with score sheets on which they are to make a judgment as to whether the second tone in each pair was higher or lower in pitch than the first of the pair. The first series of ten pairs of tones varies by 30 vibrations from the standard which is about A♯ (435 vs.). This difference is apparent to almost everybody of school age or above. The sixth series of ten tones differ by only ½ vibration and this difference is too small to be detected by any but the very best. The other differences vary between these extremes. There are four other records: one for Intensity (loudness) Discrimination; another for Sense of Time; another for Consonance (harmony) and the last for Tonal Memory.[1] These five elements Professor Seashore after much experimentation found to be fundamental in music. The last four tests are arranged very much like the pitch test. There are one hundred judgments to be made in all but the last two tests and fifty in each of these. Norms are given for fifth and eighth grade children and for adults.

Scoring the Tests—Keys are provided for scoring each of the tests. The scores are given in terms of the per cent of right responses. For example, in the test for pitch, if ten errors were made the score would be 90. By means of a table this per cent right score is transferred into a rank score. This rank score represents the standing a person would have among 100 unselected individ-

[1] Prof. Seashore includes a sixth test, musical imagination. There is no phonograph record for this test and it is not ordinarily included with the others.

uals of his group on the test. An adult making a score of 90 on the Pitch test would have a rank score of 96, that is, there would be only four better than he, in an ordinary group of 100 adults, in pitch discrimination. If it were an eighth grade child that made a score of 90, his rank standing would be 98. The same method of scoring and ranking is used with each of the other four tests except that a per cent score of 90 in any of the other tests does not give the same rank as in the Pitch test. For example, a per cent score of 90 gives a rank score of 58 for adults on the Intensity test, that is, there would be 42 better than this person in an ordinary group of one hundred adults. The keys, table for transposing per cent scores into rank scores as well as a description of the methods of giving and interpreting the score, are given in a Manual of Instruction and Interpretation for Measures of Musical Talent which is furnished by the Columbia Graphophone Company of New York City.

Function of the Tests—The Seashore Music Tests were one of the first successful vocational guidance tests to be constructed. By two or three hours' use of this test with a group of fifth grade children very reliable measures of probable success in music can be obtained. The aim of the test is not so much to measure present attainment as future possibilities in music. The tests give an answer to the question as to whether it is worth while to give the child a musical education and the amount of success to be expected. If the child ranks 90 or above on each of the tests, he is most likely to succeed in music. If he ranks below 70 in the tests, he may be given the usual school training in music but it would be unwise to spend much time or money on a musical education unless there

were other very strong reasons for thinking the child would succeed.

One of the strongest of these reasons is whether the child is vitally interested in music. If so, he is likely to make at least some success with mediocre ability. If he is not interested he is not likely to succeed with even superior native ability. The author recommends a re-survey of the pupils in the eighth grade as a check on the earlier test. The tests may also be used with high school pupils and adults. As a rule they have less practical value here as music is generally begun, if at all, earlier than this in life. The tests may also be used to advantage as an entrance requirement to college courses in music.

The Courtis Standard Supervisory Tests in Music

Description of the Tests—There are two parts to these tests; one part measures ability to recognize characteristic rhythms, and the other the recognition of mood from melody. Parts of ten Victor records are played as material for the test, five for each part. A story is first told the class and then part of a selection is played to illustrate one of four possible answers to a question in the story. The four possible answers are given on the test blank and the pupil is to mark the right answer as he interprets it from the music. There are five different sets of questions in the first part dealing with what John or someone else did and five in the last part describing how John or someone else felt about certain situations.

Uses of the Tests—These are tests in the appreciation of music for use in grades 4 to 12. They are to be used as a class test. Blanks are furnished the pupils with the

COURTIS STANDARD RESEARCH TESTS

Test 1.—Recognition of Characteristic Rhythms

Rhythm is one of the main elements of music. It has been defined as measured motion. Rhythmic motion also occurs in many of the activities of life. In this test you will be asked to judge from the music played, what life activity is represented.

JOHN'S HOLIDAY

1. It was the first day of the vacation. John had decided to go to a nearby city for a holiday. The music will tell you how John made the journey. Underline the words which tell how the music says he traveled.

1. On foot.	3. On skates.
2. By boat.	4. On horseback.

Follow this story by the Introduction and first eight measures of the Barcarolle—Victor Record No. 17311.

Test 2.—Recognition of Mood from Melody

Melody is the expression of a thought in music. In this test you will be asked to judge from the music played what John's thoughts were.

5. John's time was up now, so he took his pail and started for home. Listen to the selection and underline the words which best express how the music says John felt when his mother looked at what he had.

1. He was *sorry* he had been cross about going.
2. He was *glad* he had so many berries.
3. He was *ashamed* that he had so few berries.
4. He was *disappointed* that she said nothing.

Follow this story by the melody twice from the beginning of the Serenade Melancholique—Victor Record No. 6155.

story and questions for use in recording their responses. A key is provided as the basis for scoring the responses and a table for weighting the scores. These weights are so arranged that 1,000 represents a perfect score in the test. In order to provide for retests two alternative forms of the test are available.

Function of the Tests—These tests illustrate an interesting extension in the field of standard tests. Musical appreciation is not often looked upon as measurable. While the method used seems justifiable there might be some improvement in the technique of their construction. A larger number of judgments would reduce the chances of guessing the right answer. Before the tests are practicable for ordinary school use the selections must be reproduced on a few records specially prepared for the purpose.

Materials Needed

Courtis, S. A., The Courtis Standard Supervision Tests in Music for grades 4 to 12. S. A. Courtis, 1807 East Grand Blvd., Detroit, Mich. Price complete material for testing 40 pupils (except for phonograph records).

Seashore, C. E., The Measures of Musical Talent (Norms given for 5th and 8th grades and for adults). Five Columbia phonograph records (to be used on any standard machine) with Manual of Directions. Price $7.50.

Selected References

Seashore, C. E., *The Psychology of Musical Talent*, Silver Burdette and Co., New York City.

CHAPTER XIII

SECONDARY SCHOOL MATHEMATICS

The Rogers Test of Mathematical Ability

Description of the Test—This test is designed to measure the "mathematical intelligence" of pupils who have had five months of formal algebra and no formal geometry. It is to be used in the first year of senior high school or the third year of junior high school as a measure of probable success in more advanced courses in mathematics. There are six parts to the test: (1) a Geometry test, (2) Algebraic Computation test (test 1 and 2), (3) Interpolation test (test 1 and 2), (4) Superposition test (test 1 and 2), (5) Trabue Language Scales (L and J), and (6) Mixed Relations test.

Before each test is given, a careful explanation of the processes involved in the problems of the test is made by the tester. The explanations for each test are outlined in the manual of directions. Eight minutes is allowed for explanation in the geometry test and twenty-two minutes for the test. This test contains six problems. In each problem certain statements are given and from these the pupils are to answer a question and then state the reason or give the proof. All the reasons are to be chosen from a series of facts presented on the pages opposite the problems.

One-fourth of a minute is given for explanation of the

algebraic computation tests and three minutes allowed for the first test and seven minutes for the second. There are eleven simple problems in the first test and seven more difficult problems in test 2.

A similar method of distributing the time between explanation and work on the test problems is used in each of the other tests. The Interpolation test consists in supplying certain missing figures in a number series. For example, series are given such as the following:

A.	1	3	5	7	..	11	13	15	17	..	21
D.	1	8	15	..	29	36	43	..	57	64	71
R.	11	..	..	..	..	66	..	..	..	..	121

in which the missing numbers are to be filled in to complete the series. The Superposition test consists in locating the position of a circle in the corner of a given parallelogram by revolving a similar figure in imagination to fit on a given base line. The language test contains two parts of the Trabue language completion test. The Mixed Relations test is a statement of a proportion. The first two terms consist of words related in some way, with a fourth term to be filled in to bear the same relation to the third as the first does to the second.

Function of the Test—"The Rogers test represents six measuring rods of abilities, which after an intensive study of the activities demanded by high school mathematics were selected as possessing the highest predictive power."[1] If as the result of about an hour's time spent on such a test as this we are able to predict with reasonable assurance the possible success of a high school pupil or class

[1] *The Rogers Test of Mathematical Ability, Manual of Directions,* Teachers College, Columbia University, p. 3.

in mathematics, we have performed a real service. In such measures as these lay the future success of a large part of the real work of vocational guidance.

Algebra

The construction of standard tests in Algebra presents practically the same problems as the construction of tests in arithmetic. The selection and construction of tests in the fundamental operations is relatively simple. The application of these principles to problem solving presents a more difficult task in test construction. Yet the subject matter of courses in algebra is fairly well standardized and test construction is not so difficult as in such subjects as history or geography.

The Hotz First Year Algebra Scales

Description of the Scales—These scales, devised by Prof. Henry G. Hotz of the University of Arkansas, consist of a series of five sets of exercises. The first set consists of exercises in (1) Addition and Subtraction, the second of exercises in (2) Multiplication and Division, the third in (3) Equation and Formula, the fourth of (4) Graphs and the fifth of (5) Problems. The first two sets are designed "to test the achievement of students in the fundamental operations involving integral fractions and radical expressions; the second two to test the ability of students in handling the instruments of quantitative thinking; while the last is composed of verbal problems of the type usually stressed in first year algebra.[2] The exercises in each set are arranged in order of difficulty.

[2] H. G. Hotz, *Teachers' Manual for First Year Algebra Scale*, Teachers College Publication, Columbia University.

The first problem of each test is so easy that it may be solved by practically every pupil in the class. Each succeeding exercise is more difficult than the preceding.

There are two series of scales. Series B is the longest and contains from eleven to twenty-five exercises in each set of problems. Series A is only about half so long and contains from eight to twelve exercises in each set. In this scale the difference in difficulty is approximately the same throughout the exercise in all five sets of problems, that is, problem No. 8 is as much more difficult than No. 7 as No. 3 is more difficult than No. 2. Series A is recommended for ordinary school room use as it gives a satisfactory ranking of pupils and requires only about two school periods. When even this amount of time is not available the Equation and Formula and Problem exercises may be given in one class period. Series B is recommended especially for diagnosing the difficulties of an individual or a class.

Giving and Scoring the Scales—It is recommended that if the whole scale is to be given, the tests should be used in rotation somewhat as follows: At the end of three months, Addition and Subtraction, Equation and Formula. At the end of six months, Multiplication and Division problems. At the end of nine months, Equation and Formula (repeated), Graphs. The time allowances are twenty minutes for each of the first three sets of exercises and twenty-five minutes each for the last two in series A. In series B forty minutes is allowed for each exercise except for graphs and twenty-five minutes is allowed for it.

The scoring is made simple by entirely neglecting the

principles of solution and accepting correct answers only. Answers which may be accepted as correct are given in the manual.

Function of the Scales—The Hotz Scales fulfil the requirements of a good standard scale. The subject matter is based upon a careful analysis of the material that should be taught in the first year course in algebra. The material is arranged in each set of exercises in scale form on the basis of difficulty. The time required for the tests is not excessive, the scoring is definite and objective, and reliable norms are given. About the only serious objection that might be raised to the scales is the method of scoring only on the basis of right answers and there is at least the justification of simplicity and objectivity in this.

TENTATIVE MEDIAN STANDARDS OF ACHIEVEMENT

Series A

	3 mos.		6 mos.	9 mos.
Addition & Subtraction....	5.0		6.8	7.9
Multiplication & Division..	5.3		6.3	7.9
Equation & Formula.....	4.9		7.1	7.8
Problems	4.3		4.9	5.6
Graphs		2.8 (4½ mos.)		5.6

Series B

	3 mos.		6 mos.	9 mos.
Addition & Subtraction....	9.7		12.9	14.4
Multiplication & Division..	9.6		14.0	16.3
Equation & Formula.....	7.8		14.3	16.0
Problems	5.4		6.5	7.5
Graphs		3.7 (4½ mos.)		7.2

The Douglass Standard Diagnostic Tests for Elementary Algebra

Description of the Tests—These tests, devised by Prof. H. R. Douglass of the University of Oregon, consist of a series of forty problems. As the result of a questionnaire sent to teachers of mathematics in high schools and colleges throughout the country, ten exercises were constructed to cover the widest possible range of abilities in each of the four operations in algebra ranked as most fundamental by these teachers. These four fundamental operations are: (1) collection of terms (addition and subtraction), (2) multiplication, (3) division, and (4) solution of simple equations.

In order to get as wide a distribution of types of problems as possible, fifteen standard texts in algebra were studied with this in view. The problems were selected from the different texts, and in order to avoid any effect of practice each exercise was slightly changed from its original form.

Giving and Scoring the Tests—Each of the four sets of exercises is printed separately. Although the tests are power tests rather than speed tests—that is, the pupils are given sufficient time to solve most of the problems which they have the ability to solve—a time limit is placed on each test. Five minutes is allowed for the Collection of Terms, seven minutes for Multiplication, nine minutes for Division and eight minutes for solution of Simple Equations. The problems in each test are weighted [3] and a pupil's score in a test is the sum of the

[3] The weighting was determined by a rather complicated mathematical procedure in terms of Probable Error. See Manual, pp. 30-31.

weights for the exercises solved correctly. Average weighted scores as determined from 1,000 scores for the month of February are given as follows:

Test I 16.23
Test II 24.03
Test III 16.91
Test IV 20.93

Function of the Tests—The general principle of construction of these tests is very similar to the Hotz Scales. Douglass describes in the Manual the method used in selecting the operations to be tested. In his selection he includes four of the five operations of the Hotz Scales. Hotz adds a set of exercises in graphs. The Douglass Tests require much less time. The whole series of tests can easily be given in one class period. If the results are as reliable as the Hotz scores, this shorter time is a decided advantage in favor of the Douglass Tests. The different methods of taking care of the relative difficulty of the exercises are both equally good. A comparative study of the two tests would be most interesting. Until such a study is made the teacher must decide which fits her general purposes best. They both have the characteristics of good standard tests.

The Illinois Standardized Algebra Tests

Description of the Tests—There are four different tests in this series. The first test consists of twenty simple problems in addition and subtraction. The problems of test II call for a knowledge of transposing, test III for the removal of the parenthesis and test IV for the reduction of fractions. Four minutes is allowed for the first test,

Exercises from Each of the Four Tests of the Illinois Standardized Algebra Tests

TEST I

1. $5x - 7 = 3x - 15$
2. $-7x + 15 = 5x - 57$
3. $13x + 16 = -9x - 19$
4. $17x - 23 = -11x + 65$
5. $8x - 9 = 11x - 3$

TEST II

1. $13x - 6x = 70 - 14$
2. $-11x + 7x = 45 - 25$
3. $9x + 4x = -41 - 30$
4. $13x - 7x = -23 + 17$
5. $5x - 17x = 35 - 83$

TEST III

1. $-4(11x - 7) = 33x - 126$
2. $3(-3x + 4) = 7x - 100$
3. $-7(3x + 9) = -6x - 13$
4. $9(7x - 19) = -42x + 354$
5. $-7(3x - 5) = 14x - 7$

TEST IV

1. $\frac{3x - 14}{4} = \frac{8x - 15}{6}$
2. $\frac{-(-5x + 3)}{4} = \frac{8x - 7}{3}$
3. $\frac{-(3x + 5)}{7} = \frac{-(-5x - 3)}{4}$
4. $\frac{5x - 4}{7} = \frac{-(-3x + 9)}{5}$
5. $\frac{11x - 4}{7} = \frac{13x - 3}{8}$

five minutes for the second, nine minutes for the third and ten minutes for the fourth. Tentative norms are given as follows:

	FIRST SEMESTER		SECOND SEMESTER		THIRD SEMESTER	
	No. problems attempted	*No. right*	*No. problems attempted*	*No. right*	*No. problems attempted*	*No. right*
Test I.....	9.8	5.0	10.5	6.4	11.7	8.7
" II.....	10.8	4.6	11.8	6.4	12.6	8.3
" III....	11.0	3.6	12.2	5.5	13.9	7.3
" IV....	8.8	1.0	11.3	3.8	13.2	5.7

Function of the Tests—These tests are satisfactory general survey tests in algebra. The subject matter of these tests is based upon a study of the processes most used by pupils in solving exercises in algebra. The tests are partly diagnostic but do not test as wide a range of ability as the Hotz and Douglass tests.

The Kelley Mathematical Values Test Alpha

Professor Truman L. Kelley made a very careful study of the value to be derived by the study of algebra [4] as a result of a questionnaire to teachers of mathematics and another to a group of "capable and successful Americans." He then constructed a test consisting of thirty-eight problems to measure the thirteen fundamental mathematical values most often cited. Mathematical Values Test Alpha is partly a test in general mathematics and partly in algebra. It differs greatly from traditional tests in that the problems cover a broader field than the usual examination in algebra. Although the test may

[4] Truman L. Kelley, "Values in High School Algebra and Their Measurement," *Teachers College Record*, Vol. XXI, No. 3, May, 1920.

not have great value, it is worth the time of any teacher or parent to study Kelley's analysis of the general values of mathematics.

The Rugg-Clark Standardized Test in First Year Algebra

Any account of algebra tests, however brief, would not be complete without mention of the Rugg-Clark Standardized tests in First Year Algebra. These were the first of the better known tests in the subject and have been used extensively. These tests are no longer available but have been supplanted by the Rugg-Clark Practice Exercises in Algebra.

The Minnick Geometry Tests

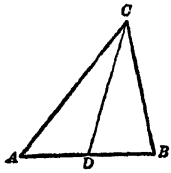
Description of the Tests—The Minnick Geometry Tests are composed of four different tests. The first, test A, is a test in construction. The exercises are given and the pupil is required to draw the figure. There are five exercises in test A. In test B the pupil is to state what is given and what is to be proved in each given exercise. The exercises are not to be solved. There are four exercises in this test. In test C a figure with certain facts about it is given, and the pupil is called upon to state as many more facts about the figure as he can. There are four exercises in this test. In test D a figure is given, the exercise is stated, other known facts are given, the problem is stated and the pupil is required to write the proof. There are three exercises in this test. Thirty minutes is allowed for each of the four tests.

Scoring the Tests—A key is provided for aid in scoring. Both a positive and negative score are to be kept. The

FIG. 14. EXERCISE FROM EACH OF THE MINNICK GEOMETRY TESTS

I. Draw the figure for the following proposition:

If two radii of a circle are perpendicular, and a tangent to the circle cuts these radii produced at points A and B, the other tangents drawn from A and B are parallel.

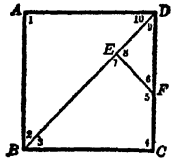


IV. State what is given and what is to be proved in the following proposition:

An angle of a triangle is a right angle, an acute angle, or an obtuse angle, according as the median drawn from the vertex of the angle is equal to, greater than, or less than one-half of the opposite side.

Given:

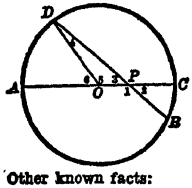
To Prove:



II.

GIVEN: The square ABCD, the diagonal BD, EB = CD and EF is perpendicular to BD.

State as many more facts about this figure as you can.



II.

GIVEN: P is any point within the circle O, AC is a diameter through P, BD is any other chord through P, OD is a radius.

To prove that AP > DP.

Other known facts:

$\angle 6 = \angle 3 + \angle 4$.

PD − OD < OP.

OD + OP > DP.

$\angle 2 = \angle 3$.

AO = OD.

$\angle 5 + \angle 6 = 180°$.

AO + OP = AP.

Proof:

positive score is based upon the number of correct statements made in each exercise. These correct statements are given with a certain value assigned on the scoring key. The per cents of correct statements are calculated and certain given values for each exercise are to be multiplied by this per cent value. The positive test score is the sum of these products. For example, in test A for exercise IV, there are 15 points in the construction of the figure. If a pupil gets 10 of these correct his per cent correct is 66⅔. The value assigned to this exercise is 23. The pupil's score for this exercise is, therefore, $23 \times 66\frac{2}{3} = 15\frac{1}{3}$. The negative score is found in the same way for all incorrect statements made in answering the exercise. Standard scores for each test are as follows:

STANDARD SCORES FOR THE MINNICK GEOMETRY TESTS

	Positive	*Negative*
Test A	62.5	7.1
Test B	69.3	3.5
Test C	50.6	4.1
Test D	73.3	2.6

Function of the Tests—These four tests not only measure ability in geometry but also indicate where the pupil is weak and where he is strong in the subject, that is, the tests are diagnostic. The time required for the tests unfortunately is too long. This could be remedied by shortening each test so that two tests could be given in one high school period. A simpler method of scoring would make the tests more usable by the average classroom teacher. These are problems in the administration of the tests. The fundamental basis upon which the tests were constructed is sound and they are without doubt the most satisfactory tests yet devised in the subject.

Materials Needed

Douglass, H. R., Diagnostic tests for First Year Algebra. H. R. Douglass, University of Oregon, Eugene, Oregon.

Hotz, H. G., Algebra Scales for classes in High School Algebra. Teachers College, Columbia University. Manual of Directions 15 cents. Sample set of scales 12 cents. Each of first four scales $0.75 per hundred. Graph scale $1.25 per hundred.

Kelley, T. L., Mathematical Values Test. Teachers College, Columbia University. One set of scales 40 cents. Teachers College Record, May 1920, 40 cents. Test blanks 5 cents each.

Minnick, J. H., Geometry tests. Four separate tests. Public School Publishing Co., Bloomington, Ill. Sample set 12 cents. Tests A, B, C and D, each $2.50 per hundred.

Rogers, Agnes L., Tests for Diagnosing Mathematical Ability for the ninth grade. Teachers College, Columbia University. Manual of Directions 50 cents. Sample copy of test 10 cents. Price $7.00 per hundred.

Rugg, H. O. and Clark, J. R., Standardized Practice Exercises in First Year Algebra. H. O. Rugg, The Lincoln School of Teachers College, New York City

Selected References

Douglass, H. R., *The Derivation and Standardization of a Series of Diagnostic Tests for the Fundamentals of First Year Algebra,* University of Oregon, 1921.

Hotz, H. G., *First Year Algebra Scales,* Teachers College, Columbia University, 1918.

Kelley, T. L., "Values in High School Algebra and Their Measurement," *Teachers College Record,* May 1920.

Minnick, J. H., *An Investigation of Certain Abilities Fundamental to the Study of Geometry,* University of Pennsylvania, 1918.

Rogers, A. L., *Experimental Tests of Mathematical Ability and Their Prognosis Value,* Teachers College, Columbia University, 1919

CHAPTER XIV

SECONDARY SCHOOL SCIENCE

At first thought one might expect that standard tests in the sciences would be among the easiest to construct, but this is not the case. Several causes contribute to the difficulty of such tests. First, there is less agreement as to the content or subject matter of courses in science than in most of the other school subjects. This is due partly to the fact that there is such a large body of material from which to select and partly to differences in the aims of teaching these subjects. Some teachers emphasize the factual or informational side of the sciences, some the laboratory practice, some the thought or reasoning elements, while some have still other aims, no aims or a combination of many aims. For these reasons it is difficult to determine what material to include in the tests. A further difficulty arises in the relative weights to be given to the different parts of a test. If tests in chemistry are to include numerical or laboratory problems, how many and what weight should they receive? Such tests as the Downing Test in Science, are almost entirely factual or a test of information. Other tests, like the Rich Chemistry Test, attempt to distribute the problems so as to include questions in thinking, memory, numerical computation and laboratory exercises.

Descriptions of only the most representative scales and

tests in General Science, Physics and Chemistry are included in this chapter. These include the Downing Range of Information Test in Science, the Van Wagenen Reading Scales in General Science, the Rich Chemistry Tests and the Iowa Physics Test.[1]

The Van Wagenen Reading Scales—General Science Scale A

Description of the Scales—The form of these General Science scales is very similar to the other Van Wagenen scales already described in the chapters on History and Geography. There are short paragraphs on various topics in general science, followed by sets of four to six questions based on the subject matter of the paragraphs. There are fifteen paragraphs in all, divided into three groups of five paragraphs each. The pupil is directed to read each paragraph carefully. "Then read the statements below it and put a check mark (√) on the dotted line in front of each statement which contains an idea that is in the paragraph or that can be derived from it."

Scoring the Scales—A key is provided for use in scoring the tests. The number of errors are counted separately for each of the three groups of paragraphs. The uncorrected score for a pupil is obtained from the key on the basis of the number of errors made in the questions to the paragraphs of group III. This first score is corrected in relation to the number of errors made in the answers to the questions in group II and recorrected likewise for errors in group I to give the final score.

[1] A more complete list of tests in science as well as in the other school subjects may be found in the little pamphlet entitled "Bibliography of Tests for Use in Schools," World Book Co., Yonkers, N. Y. Price 10 cents.

KEYS FOR USE IN SCORING THE VAN WAGENEN READING SCALE—GENERAL SCIENCE A

KEY A

Paragraph number	Statements to be checked
Group I	
1	1, 4
2	2, 5
3	1, 2, 3
4	3, 4
5	3, 4
Group II	
6	2, 4
7	2, 3, 5
8	1, 3, 4
9	3
10	2, 3, 4
Group III	
11	2, 5
12	2, 5
13	2, 3, 5
14	2, 5
15	2, 3, 4

KEY B

(For use in obtaining the Uncorrected Score)

When errors in Group III are	1	2	3	4	5	6	7	8	9	10
Pupil's Uncorrected Score is	107	100	97	94	92	90	88	86	84	83

KEY C

(For use in obtaining the Final Score)

When errors in Group I are....	0	1	2	3	4	5	6	7	8	9
Take from the First Corrected Score	0	0	1	2	3	4	5	7	10	15

Tentative norms are given as follows:

TENTATIVE STANDARDS OF ACHIEVEMENT

For the pupil who stands at the	Grade 8	Freshman	Sophomore	Junior	Senior
25-percentile (¼ way up from the lowest)	72	74	77	80	83
Median (½ way up)	77	80	84	87	90
75-percentile (¾ way up)	83	85	89	92	97

Meaning of the Pupil's Score [2]—The scores yielded by this test have no relation to per cents. A score of 71 means that the pupil who makes it can read paragraphs and sets of statements of difficulty 71 and get one-half of

[2] From the Class Record Sheet.

them correct, or its equivalent. It also indicates that the pupil can read paragraphs and sets of statements of difficulty 61 and get three-quarters of them right. At the same time, if the pupil were given paragraphs and sets of statements of value 81, he would be most likely to get one-quarter of them correct. Throughout the scale the difference between any two points is equal to a similar distance between any other two points. For instance, the pupil who gets 81 is doing just as much better than the pupil who gets 71 as the pupil who gets 71 is doing better than the pupil who gets 61.

Function of the Scales—These are very satisfactory scales for measuring the combined product of information and ability to gain information from reading in general science. If the subject matter of the paragraphs were new material to the pupils, it would then become a test of ability to gain information by reading. Since the material is more or less familiar to the pupils, some of the questions are likely to be answered from the pupil's stock of information rather than from his careful reading of the paragraphs. In many cases these tests exactly meet the needs of the teacher. If the teacher wishes to measure range of information alone some other test is preferable. The administration of the test would be made easier by a more simplified method of scoring.

The Downing Range of Information Test in Science and the Grier Range of Information Test

Description of the Tests—The test devised by Dr. Elliot R. Downing consists of a list of fifty words or phrases selected from the various sciences. The terms are well distributed between physiology, geography, biology,

physics and chemistry. The words are arranged alphabetically on the test sheet. The pupil is directed to put an "E" beside the words and phrases that he can explain or define, and "F" beside the ones he has heard or read about, the meaning of which is not clear, and an "N" beside those that are new. He is then directed to explain or define the first five marked with an "E." The pupil is given all the time needed to take the test. The answers are scored on the basis of the number of words marked in each of the three groups "E", "F", and "N", except that the "E" list is reduced by the percent of words wrongly defined and the reduction added to the "F" list. For example, if a pupil has 15 words marked with an "E" and 15 with an "F" and 10 with "N", and one of his definitions is wrong, his score should be E-12, F-18, N-10. The norms for the ninth grade are given as:

THE DOWNING RANGE OF INFORMATION TEST IN SCIENCE NORMS

"E"..... 18.6 "F"..... 13.4 "N"..... 18

Section from the Revised Range of Information Test in Science

By Dr. Elliot R. Downing

Please put an E beside words and phrases (on the list below) that you can explain or define, an F beside those you have heard or read about, the meaning of which is not clear, and an N beside those that are new. Explain or define the first five you mark with an E, on the back of this sheet.

No.	*Mark Here*		No.	*Mark Here*	
1		Adaptation	26		Inoculation
2		Atom	27		Instinct
3		Buoyancy	28		Law of gravitation
4		Candle power	29		Law of the lever
5		Center of gravity	30		Law of the pulley

The Grier Range of Information Test is very similar to the Downing Test, except that there are three lists of one hundred words each. One list is made up of words selected from physiology, another from zoölogy and another from botany.

Function of the Tests—Either of these tests is very satisfactory as a method of rapid survey of general information in the fields of general science. They are factual tests. There is no attempt to weight the relative importance of the different items of the test nor any attempt to measure the pupil's ability to use his information. They are survey tests and in no sense diagnostic. Despite these limitations the tests provide the teacher with a ready measure of her pupils' fund of information in general science.

The Ruch-Popenoe General Science Test

Description of the Test—"This test is designed primarily to measure the accomplishment of pupils in general and elementary science courses in either the eighth or the ninth grade. It is not based upon any single textbook in general science nor is it intended to apply to any particular type of a course of study in this subject."

"Part I is composed of fifty items of general information concerning familiar elementary scientific facts, principles, concepts, terms, definitions, and applications. These fifty items sample a wide range of relatively simple knowledge in the fields of physics, chemistry, astronomy, agriculture, botany, zoölogy, and physiology."

"Part II measures the ability of the pupil to identify apparatus, organisms, structures, and principles, and to apply principles of science to the solution of simple prob-

Ruch-Popenoe. Form A

24 The length of a meter in inches is about 12 19 24 2 39 47 144..... 24

25 A general term for any living thing is a
plant cell larva animal organism mammal nucleus..... 25

26 A violent circular windstorm of small area is a
cyclone, tornado monsoon trade wind norther blizzard equinox..... 26

27 A collection of similar cells is called an
organism tissue organ gland muscle sense-organ function 27

28 The watt is the unit of measurement of
resistance current velocity power potential inductance friction..... 28

29 The unborn young of an animal is termed the
larva embryo pupa adult chrysalis ovum sperm..... 29

30 An example of a chemical element is
water glass mercury carbon dioxide ammonia nitric acid sugar..... 30

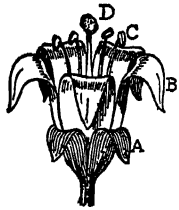


FIGURE 2

In this diagram of a typical flower:

a The petals (the corolla) are marked by the letter..... *a*

b The stamens are marked by the letter ..... *b*

c The sepals (the calyx) are marked by the letter ..... *c*

d The pistil is marked by the letter ..... *d*

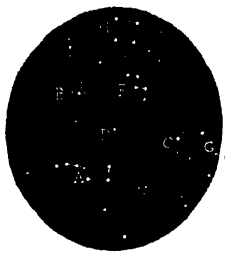


FIGURE 3

a The North Star (Polaris) is marked by the letter..... *a*

b The Big Dipper is marked by the letter ..... *b*

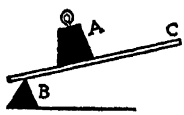


FIGURE 4

a In this lever the power or force is applied at ..... *a*

b The fulcrum is placed at the point marked ..... *b*

c The mechanical advantage of a lever of this class is always .....than 1. *c*

[4]

FIG. 15. SAMPLE EXERCISES FROM THE RUCH-POPENOE GENERAL SCIENCE TEST

lems. For this purpose twenty diagrams and drawings have been prepared, together with completion exercises based upon these illustrations."

The distribution of subject matter among the different subjects is given as follows:[3]

SUBJECT-MATTER ANALYSIS OF THE RUCH-POPENOE GENERAL SCIENCE TEST IN PERCENTAGES

Biological science (botany, physiology and zoölogy)	30%
Chemistry	12%
Physics and mechanical applications	38%
Earth science (agriculture, astronomy, geology, and physiography)	20%

Fifteen minutes is allowed for Part I and twenty-five minutes for Part II. There are two alternative forms of the test. Norms are given for the eighth and ninth grades as follows:

TENTATIVE NORMS FOR THE RUCH-POPENOE GENERAL SCIENCE TESTS

8th Grade	34.0
9th Grade	41.0

Function of the Test—This is without doubt more comprehensive than either of the other tests in General Science. Part I takes the familiar form of general information tests. It is extensive enough to give a reliable measure of general information in science. Part II is unique and measures a different type of ability. It would be interesting to determine how successfully actual laboratory conditions are duplicated by the drawings and how closely ability in this part of the test correlates with laboratory knowledge and technique. The test is easy to administer and full directions are given for tabulating scores on a percentile or ogive curve.

[3] Ruch-Popenoe General Science Test, Manual of Directions, p. 1.

The Rich Chemistry Tests

Description of the Tests—These tests consist of twenty-five questions in chemistry. "The cyclic principle of Rugg is applied in test construction using a cycle of six questions: thinking, memory, numerical, thinking, memory, laboratory. Within each kind of question there is a steady increase in difficulty from those in the earlier part of the tests to those in the later, and for the questions as a whole the same is approximately true. The subject matter of the tests was derived from material common to at least two widely used texts in chemistry or to one text and questions recently given in the College Entrance Board Examinations, the New York State Regents Examinations, or from material found in state syllabi in chemistry. Twenty-five minutes is allowed for the tests. There are at present two forms of the test and other forms are promised later.

Scoring the Tests—Most of the questions are answered by checking the right answer from a list of four possible answers. In the numerical type of questions the calculation is to be done in a space on the test blank provided for the purpose. The score is given in terms of a T-score. This T-score is determined on the basis of the number of questions answered correctly and is obtained from a table given in the Manual. As the author explains, these T-scores are directly comparable from one test to another.[4] Norms are given by half semesters for the work of the first two years for both high school pupils and college students.

[4] It is also explained that T-scores are not per cents and a further table is provided for turning T-scores into per cent scores or letter grades on the basis of 70 as the lowest passing grade.

NORMS FOR THE RICH CHEMISTRY TESTS

Time pupils have studied chemistry	*High School Pupils*	*College Students*	*All*
.5 Semester	43.6	45.5	45.0
1.0 "	44.5	...	45.0
1.5 "	48.8	51.2	49.1
2.0 Semesters	53.3	54.4	53.4
2.5 "	53.3	55.2	54.3
3.0 "	...	60.3	58.4
3.5 "	56.2	57.6	58.5
4.0 "	...	60.3	...

... means insufficient scores to establish reliable norms.

Function of the Tests—These tests furnish the teacher of chemistry with a ready means of checking up on the pupil's work in that subject. The tests are carefully constructed so as to test the various materials of instruction; they are almost wholly objective and are easily administered and scored. Since the tests are spirally arranged it is possible to determine in which of the five types of questions the pupils are strong and in which types they are weak. For example, questions 2, 5, 8, 11, etc., are memory questions; 3, 9, 15, 21 and 24 numerical questions; 6, 12 and 18 laboratory questions. By inspection of the scores of a pupil or a class, it is possible to determine the strengths and weaknesses of the pupils. This information in turn can well be made use of in determining what should receive special emphasis in later instruction. In other words, these tests have a high diagnostic value. On the whole these are among the most satisfactory tests in the content subjects of the secondary school curriculum.

The Iowa Physics Test

Description of the Test—This test was devised by Dr. H. L. Camp while a graduate student at the University of Iowa. It consists of three sets of problems: one in mechanics, one in heat, and one in electricity and magnetism. There are twelve problems in the first set and eleven in each of the other two sets. The questions are largely factual, covering a representative as well as fairly wide range of information in each of the three fields of elementary physics. The answers are brief and are to be written in the spaces following each question. The tests in mechanics and heat take forty-five minutes and the test in electricity and magnetism takes forty minutes.

Scoring the Test—Each problem is given a weighted value and a pupil's score is the sum of the values of the questions answered correctly. The following are tentative median scores:

TENTATIVE MEDIAN SCORES FOR THE IOWA PHYSICS TEST

	Mechanics	*Heat*	*Electricity & Magnetism*
Boys	39.5	44.4	46.3
Girls	30.6	40.1	36.5
Total	33.5	41.9	39.9

Function of the Test—This test meets the need of a measure of information in each of the three main subdivisions of physics. The subject matter is fairly extensive and representative, the scoring is brief and simple. The test fails to present a carefully selected and systematically arranged series of problems as the Rich Chemistry Test. It also lacks such extensive diagnostic value.

Two Questions from the Iowa Physics Test Electricity and Magnetism

Series C. Form 1. By Dr. Harold L. Camp

Value	
(9.4)	6. What property of a volt-meter prevents it from short-circuiting the two lines when connected across? **ANSWER**
(10.2)	7. What property of an electric current is utilized in comparing currents by means of a galvanometer? **ANSWER**

Materials Needed

Camp, H. L., The Iowa Physics Tests for Secondary Schools. Separate tests on Heat, Mechanics and Electricity and Magnetism. Packages of each. Price 50 cents per package of 25. Public School Publishing Co., Bloomington, Ill.

Downing, E. R., Range of Information Test in Science for grade 9. Directions and scoring sheet 10 cents. Test sheets 40 cents per hundred. E. R. Downing, University of Chicago, Chicago, Ill.

Rich, S. G., Chemistry Tests, Gamma and Epsilon for use in high school and college. Sample set 20 cents. Teachers Manual 15 cents. Price of tests per package of 25, $1.00.

Ruch, G. M., and Popenoe, H. F., General Science Test for grades 7 to 9. Forms A and B. Specimen set 25 cents. Price for class of 25 complete $1.50. World Book Co., Yonkers, N. Y.

Van Wagenen, W. J., Reading Tests; General Science for grades 8 to 12. Forms A and B. Sample set 20 cents. Price $3.00 per hundred. Public School Publishing Co., Bloomington, Ill.

Selected References

Briggs, T. H., "General Science in Secondary Schools," *Teachers College Record,* Vol. XIII, pp. 19 ff.

Caldwell, O. W., *Science Teaching,* General Education Board, New York, 1919.

Judd, C. H., *The Psychology of High-School Subjects,* Chapt. XIV, Science, Ginn & Co.

Rich, S. G., Chemistry Tests, Teachers' Manual.

Trafton, G. H., *The Teaching of Science in Elementary Schools,* Houghton Mifflin Co.

CHAPTER XV

FOREIGN LANGUAGES

I. Latin

The use of objective tests as a means of determining certain specific ends to be attained in the teaching process has been used in numerous cases in a limited way, but the American Classical League is the first body of teachers to attempt the determination of teaching objectives through a nation-wide use of such instruments. In their recent investigations, they have given over one million tests in one thousand high schools. Through their work, a large number of Latin teachers have become familiarized with certain tests, and the diagnostic use to be made of them; psychologists have been stimulated to work out several series of Latin tests which might not otherwise have come into use; and Latin teachers in the future will be inclined to use these same tests in their class work, since such definite norms have been worked out for them. Accordingly, a statement will be made here of the purposes which were served by the use of the tests, and a brief description given of the instruments used.

Several objectives have been suggested as goals legitimately to be striven for in the study of Latin. Actual ability to read the language after the study in school has ceased, is the first objective suggested to the mind of the

ordinary layman. So tests were devised to determine the extent of the ability of the pupil to read the language.

Another value commonly ascribed to Latin study is the increased ability to understand and use English words derived directly or indirectly from Latin. Studies and tests have been made to determine the validity of this claim.

Again, the claim has been made that the study of Latin aids the pupil in the mastery of other subjects, especially other foreign languages, and English. Tests have been devised to make possible comparative results of Latin and non-Latin pupils in such fields of learning.

The so-called disciplinary values of Latin have also been widely discussed pro and con during past years. An attempt has been made to do some diagnostic experimental work to throw some light upon this topic of general interest.

As a result of the attempts made to measure these various objectives, the Classical League says with justifiable pride: "No other high school subject can at present show a list of standard tests comparable to ours." [1]

The tests—Three types of tests were used: (a) Tests to determine the knowledge which the pupil had in English; (b) tests to determine the pupil's background in Roman History, either antecedent to, or correlative with, his Latin study; and (c) tests in various skills in the language itself.

(*a*) *Knowledge of English*—These tests will be mentioned only. They were:

The Thorndike Special Vocabulary Test.
Thorndike-McCall Reading Scale, forms 1, 2, 3, 8.

[1] *Classical Journal*, Vol. XVII, p. 561.

Carr English Vocabulary Test, forms A, B, C, D (made up of Latin derivatives).
Buckingham-Coxe English Spelling Test, forms A, B, C.
Thorndike Test of Word Knowledge, forms A, B, C, D.
The Illinois Educational Research Special Reading Test.

This list is recommended to teachers who wish to discover the English background of their pupils, and so their probable adaptability to the study of Latin. Other tests along similar lines will be found in the chapters on Spelling, Reading, and English Composition.

(*b*) *History Background*—Two tests were used here, the Davis-Hicks True-False Test in Roman History, Late Republican Period, and the Davis-Hicks Test in Historical Content and Background of Cæsar's Gallic War. A third test closely allied with historical material, was the Clark-Ullman Test on Classical References and Allusions.

The Davis-Hicks Roman History Test is made up of fifty direct statements, followed by the words "True ... False...." The pupil is to underline either the word "True" or the word "False" according to the truth or falsity of the statement. Careful instructions for giving are printed on the front page of the test. Ten minutes is allowed for the test. If the pupil has forgotten the correct answer, but thinks the statement seems familiar, he is to put a question mark between the words "True" and "False." If the statement seems entirely new to him, he is to place a circle around the number of the question. The questions center very largely around the careers of Cæsar and Cicero, and may serve as a means of determining the information gained from the study of these authors in class, as well as a means of prognosis to de-

termine the emphasis to be placed upon the historical side in teaching the authors named.

The Test on Historical Content and Background of Cæsar's Gallic War may also be both prognostic and diagnostic. It is of the same general type as the foregoing, consisting of fifty statements regarding the War in question. Instead of the True-False arrangement, alternative statements are made, the correct statement to be underscored, thus: "1. Cæsar's conquest of Gaul probably kept it from very soon becoming British, German, Grecian, Moorish, territory." The correct word in this case is "German," so it is to be underscored. Twenty-five minutes is allowed.

The Clark-Ullman Test on Classical References and Allusions consists of fifty statements, each concluding with five possible endings, the correct one to be underscored. The statements deal with both Greek and Roman references. A sample sentence is:

> "The Greeks believed that the home of their gods was
> Athens, Mount Olympus, Vesuvius, Arcadia, Mt. Ætna."

Since Mount Olympus is correct, a line is to be drawn under that word. Twenty minutes is allowed for the test. Other samples are:

9. The mortal who stole fire from heaven was
   Prometheus, Pandora, Vulcan, Vesta, Mercury.
10. Pure water was brought to Rome by
    Tunnels, Aqueducts, Appian Way, Cloaca Maxima, Tiber River.

The value of the test to the teacher as a means of prognosis is considerable. Certainly it will give a very definite index of the child's general classical knowledge, and in

this way may well be valuable to history and literature teachers as well as to Latin instructors.

(*c*) *Latin Tests*—The tests used in the Classical investigation were designed to measure five main abilities of the pupils: (1) knowledge of vocabulary; (2) knowledeg of forms; (3) knowledge of syntax; (4) knowledge of prose composition; (5) ability to comprehend the thought of a Latin sentence or paragraph. The tests will be described in this order.

(1) *Vocabulary*—The Henmon Vocabulary Test, designed by Prof. V. C. Henmon of the University of Wisconsin, consists of fifty Latin words arranged in order of increasing difficulty. The subject is to write after each word its meaning, and each word is given a definite weight or value, so that the score is the sum of the various values of words defined correctly. The first, "bellum," is given the lowest value, .4, and the last, "quisque," is given the highest, 4.7. The steps between words vary from .1 to .3. In several cases, words are assigned equal values. This test represents seven years' study and experiment, to determine the relative weights of the words. They are taken from thirteen beginning Latin books, and are all words in frequent use in the high school Latin course. There are two forms, for alternate use, to prevent any unfair use of the test.

The Henmon Sentence Test is printed with the Vocabulary Test, and gives a test of vocabulary, in the rendering into English of ten brief sentences. These are arranged in the same way as are the vocabulary words, beginning with a three word sentence of the value 1.7, and ending with a seven word sentence of value 6.2. Sample sentences are:

2. Sunt viri fortes.
8. Equites contenderunt ut quam primum domum pervenirent.

This test also makes possible a measure of Latin comprehension, and ability to handle various types of constructions. It was devised with the same care that marks the Vocabulary Test. There are two forms of this test.

The Wisconsin Test in Latin Words, Phrases, and Abbreviations Occurring in English was devised by Lou V. Walker of the University of Wisconsin. The test is divided into three sections, the first containing twenty-five phrases of Latin origin, as "ad libitum," "e pluribus unum," "pro bono publico"; the second twenty abbreviations of Latin origin, as "ult., A.B., non seq."; and the third fifty words of Latin origin, as "memoranda," "duplex," and "stratum."

The subject is "to decide in each case the English word or phrase which means most nearly the same thing, and write the meaning after the word." Thirty minutes is allowed for the test. The test is both prognostic and diagnostic, for when given at the beginning of the Latin course, it reveals the pupil's background in such a way as to point out the method for vocabulary presentation and drill, and given during the course, makes possible an evaluation of the types of difficulty the child is meeting in his vocabulary work. The teacher will also find various hints as to the individual differences of the children in other factors than those enumerated.

Knowledge of Forms—The Tyler-Pressey Test in Latin Verb-Forms is designed as a test in forms of verbs exclusively. The authors are Caroline Tyler and S. L. Pressey of Ohio State University. The test comprises thirty-two verb-forms, each form being followed by four

translations, only one of which is correct. The subject is to underline the correct form. Thus:

"laudavit
He praises.....He praised.....He will praise.....
He had praised....."

"He praised" is correct, and so is underlined. Fifteen minutes is the allotted time. All forms are presumed to be of equal value. The test gives a thorough diagnosis of the pupil's knowledge of verbs. There are two forms of this test.

(3) *Syntax*—The Pressey Test in Latin Syntax involves nouns, pronouns and adjectives. It consists of thirty-three English sentences, each followed by four translations of that sentence into Latin. Only one of the four Latin translations is correct, and the pupil is required to underline the one which he thinks is correct. Thus:

"I see a man
Vir videō.....Virum videō.....Virō videō.....
Virus videō"

"Virum videō" is correct and so is underlined.

No values are assigned to the various sentences, and the score is based upon the number of sentences marked correctly. Twenty minutes is allowed for the test. Forms 1 and 2 are furnished if alternative tests are desired.

Syntax is closely allied with interpretation of sentences and general comprehension, and so may also be tested by the Henmon Sentence Test, and the Ullman-Kirby Comprehension Test, as well as entering into the Composition tests.

(4) *Composition*—The Godsey Diagnostic Latin Com-

position Test consists of two forms, Form 1 and Form 2. Each form is made up on the following plan: There are three sections in the test, each made up of eleven English sentences to be translated into Latin. Under each sentence is the translation, correct except for one word or phrase for which four forms are given, one only of these being correct. At the end of the line on which the sentence is written are four figures, referring to corresponding numbers of a set of rules printed at the bottom of the page. The pupil is to draw a circle around the correct form, then decide which numeral corresponds to the correct rule at the bottom of the page, and draw a circle around the proper numeral. Thus:

"The boys are in the town.
Puerī (in oppidum, oppidō, in oppidō, in oppidīs) sunt.....
.....1.....2.....7.....15."

Since "in oppidō" is correct, a circle is to be drawn around it, and since "2" is the index number of Rule 2 at the bottom of the page which says "Place in which is regularly expressed by the ablative with "*in*," another circle is to be drawn around the numeral "2."

The time allowed is thirty minutes. In this time only the most proficient pupils will be able to complete the test. Thus it is possible to use it in consecutive years as a progressive measure of attainment. Nouns, pronouns, and verbs are involved in the test. It thus involves forms, syntax, and idiomatic usage. It does not bring in vocabulary knowledge to any material extent, but it does make possible an accurate determination of the extent to which the pupil uses rules or other references intelligently in his work.

(5) *Comprehension*—Certain factors in comprehension have already been touched upon in connection with understanding of short sentences, vocabulary, syntax, and composition, but there is remaining a need for some test of connected material involving extended sentences or groups of sentences forming paragraphs. The Ullman-Kirby Latin Comprehension Test, Forms 1 and 2, was designed by B. L. Ullman and T. J. Kirby of the University of Iowa to meet this need.

It consists of ten paragraphs each containing from three to six lines of Latin. Under each paragraph is a series of three or four questions designed to test the pupil's ability to comprehend the content of the passage. He is to write a brief answer to each question, in English. The paragraph is before him for continual reference. The practice paragraph is:

> "Duōs fīliōs agricola habēbat Mārcum and Sextum. Magnus erat Sextus et Mārcus erat parvus. Pater parvum fīlium magis amābat et eum ad urbem saepe mittēbat."
>
> a. Which was more loved by his father, Sextus or Marcus?
>
> .....
>
> b. Where did his father send Marcus?.....

The passages are graded in difficulty, and comprise two poetic paragraphs. Thirty minutes is allowed for the work, but only the most advanced pupils will finish in that time. It thus allows for progressive grading. The authors believe that a better index of real comprehension is thus furnished than by an attempted translation. This test will be very suggestive to those teachers who have not hitherto paid attention to the factor of general comprehension as contrasted with mere translation.

Conclusion—The series of Latin tests here described

will be best used, if they are given as a "battery" of tests, rather than as single tests at rather wide intervals. No one test will reveal all that is necessary to be known about a class or an individual student, but taken together, they will give a decidedly accurate diagnosis of the situation. Each in a sense supplements the others. A preliminary study of the child or of the class may make wise the omission at first of a certain test or tests, but in the main, they should be used together.

II. Modern Languages

Several tests have been worked out for French and Spanish. Brief mention will be made of them, for they are applied along the same lines as are the Latin, and so the discussion already given in this chapter need not be repeated.

The Henmon French Tests are worked out along the same lines as the Henmon Latin Tests already described. They consist of four sets of tests, each comprising a vocabulary test and a French sentence test, printed together. Both words in the vocabulary, and sentences, are graded as to difficulty, and score values assigned, so that they may be scored accurately. The difficulties are so graded that the tests are not too simple for fourth year students, even. They thus may be used progressively.

The Handschin Modern-Language Tests are designed to be used by first and second year students in French and Spanish to measure silent reading, based upon quality and rate, and by French students in composition and grammar. These tests take but five minutes for the silent reading, and ten for the composition and grammar and so are to be commended as taking but little time from the

regular routine. Perhaps the teacher will feel that but little of value can be learned about the pupils in so brief a period, but the actual diagnosis resulting will be found of considerable value. As was suggested for the Latin tests, it is best to give several different modern language tests, and to evaluate the results in terms of all, rather than of any one.

Other modern language tests which have been devised are Starch's French Reading and Vocabulary Tests, Starch's German Reading and Vocabulary Tests, and Wilkins's Prognosis Test for Modern Languages. This last is designed to determine beforehand by a series of visual and aural motor and oral tests (seeing, hearing, writing and speaking) four factors which are deemed essential to success in foreign language. They are: susceptibility to impression; readiness and accuracy of expression; retentiveness of memory; and grasp of the ordinary concepts of grammar. By their use, students may be classified into slow and rapid moving groups, or may be grouped in sections in which relatively different amounts of instruction are to be given in English grammar and other drill work in connection with the beginning of the foreign language.

Materials Needed

Brown's Latin Vocabulary Test, Latin Grammar Test, Latin Sentence Test, The Parker Co., Madison, Wis.

Clark-Ullman Test on Classical References and Allusions—B. L. Ullman, State Univ. of Iowa, Iowa City, Iowa.

Davis-Hicks, "A Test on the Historical Content and Background of Cæsar's Gallic War." E. E. Hicks, Wilkinsburg, Pa.

Detroit Latin Vocabulary Test. Department of Instruction, Detroit Public Schools, Detroit, Mich.

Godsey Diagnostic Latin Prose Composition Test—Form 2 (temporarily until present stock is exhausted). Service Bureau for Classical Teachers, Teachers College, Columbia University, New York City. After stock is exhausted from Edith R. Godsey, Atchison, Kans.

Grinstead Special Latin Vocabulary Test—W. J. Grinstead, George Peabody, College for Teachers.

Henmon Latin Vocabulary and Sentence Test—The World Book Co., Yonkers, N. Y.

Inglis Latin Tests: General Vocabulary, Syntax, Morphology (Inflections). Sample set 25 cents. Harvard University Press, Cambridge, Mass.

Kansas Latin Derivation Tests A and B; Latin Teachers Tests C, D and E.—Kansas State Normal School, Emporia, Kansas.

Lohr-Latshaw Latin Form Test—Bureau of Educational Research, University of North Carolina, Chapel Hill, N. C.

McTammany Latin Speed Tests (vocabulary and prin. parts.) Frances L. McTammany, Troy High School, Troy, N. Y.

Otis English Latin Derivation Test. World Book Co., Yonkers. N. Y.

Pressey Test in Latin Syntax, Form 2 (temporarily until present stock is exhausted) from Service Bureau for Classical Teachers, T. C. Columbia Univ., afterwards from Dr. L. W. Pressey, Ohio State Univ., Columbus, Ohio.

Thorndike-McCall Reading and Scale—Bureau of Publications, T. C. Columbia University, New York City.

Thorndike Test of Word Knowledge Advanced Form—1—American Classical League, Princeton, N. J.

Tyler-Pressey Latin Verb Forms—form 2 (temporarily until present stock is exhausted) Service Bureau for Classical Teachers, T. C. Columbia University, afterwards from Caroline A. Tyler, 425 Barkley Road, Columbus, Ohio.

Ullman-Kirby Latin Comprehension Test—B. L. Ullman, State University of Iowa, Iowa City, Iowa.

White Latin Test, Part I, Vocabulary; Part II, Sentences. World Book Co., Yonkers, N. Y.

Wisconsin Test in Latin Words and Phrases occurring in English—Miss Lou V. Walker, 2814 College Avenue, Alton, Ill.

CHAPTER XVI

GENERAL ACHIEVEMENT TESTS

"The choice of standard educational tests for use in the various school grades and school subjects has for some time presented to the average superintendent or principal a problem of great difficulty. The number of such tests is legion, and each is fragmentary in the sense that it covers only a part, and often only a small part, of the ground that the average teacher or administrator desires to cover. In selecting tests for his use, the superintendent has found it necessary to make a comparative study of the many tests available for each subject and for each grade. In order to assemble a suitable battery of tests measuring the achievement of an entire school, it has been necessary to make a half-dozen or more comparative studies of this kind and to arrive at as many separate decisions. It has then been necessary, after the tests to be used have been decided on, to order each from a different publisher and to master several different types of procedure and scoring. After the tests have been assembled, given, and scored, it is often found that the norms for the various tests are expressed so differently or have been derived by such diverse methods that there is no satisfactory way to compare a pupil's score in one subject with his score in any other, or to summate a pupil's scores for the various subjects into a composite score. So extremely confusing has

the situation become that many school administrators justly hesitate to embark on a testing program at all, and the average teacher or principal is hopelessly at sea." [1]

In order to meet these problems there have been several groupings of tests into battery tests. The Illinois Examination, consisting of an intelligence test, a test in arithmetic and one in reading, was one of the first of such groupings. Other tests devised by Monroe and Buckingham, while printed on separate blanks, conform to the same general principles as the Illinois Examination and might well be considered parts of a general achievement test.

These tests along with some others have recently been more closely grouped together by expressing the results of the tests in terms of B scores. The B scores are grades in terms of standard achievement for each month of the school year. For example, a B score of 5.2 in handwriting means that the pupil has reached the norm for a pupil of the second month in the fifth grade. In this way test scores for the different school subjects are all transformed to a common basal scale and hence are directly comparable. The advantage of such a common unit of measurement is apparent. School standing and progress from school to school and from subject to subject are measured in common terms. Such methods will do much to make standard tests and scales more available for and usable by the supervisor and class room teacher.

The Pintner-Marshall [2] combined Mental-Educational Survey tests, as their name indicates, are another battery

[1] Stanford Achievement Test, *Manual of Direction*, by T. L. Kelley, G. M. Ruch and L. M. Terman, p. 3.

[2] R. Pintner and H. Marshall, A Combined Mental-Educational Survey, *Journal of Educational Psychology*, January and February, 1921.

of tests which includes a measure of general intelligence and tests in the principal school subjects from grades 3 to 8 inclusive. The Chapman Classroom Products Survey Tests contain two reading tests and two arithmetic tests for grades 5 to 8. Dr. S. L. Pressey has constructed Scales of Attainment for each of the first three school grades. The first grade scale consists of a test in reading. The second grade scale contains reading, spelling and arithmetic tests. The third grade scale contains a test in spelling, a test in rate and comprehension of silent reading and a test in the four fundamental operations of arithmetic.

The Stanford Achievement Tests

Description of the Tests—None of the general achievement tests we have thus far described are as complete or as carefully worked out and standardized as the Stanford Achievement Tests by Truman L. Kelley, G. M. Ruch and Lewis M. Terman. These tests compose a complete arrangement of tests for grades 2 to 8. The tests for grades 2 and 3 are printed in one pamphlet which is called the Primary Examination. The tests for the higher grades are printed in another pamphlet called the Advanced Examination.

In the primary examination there are three tests in reading: the first in paragraph meaning, the second in sentence meaning and the third in word meaning. There are two tests in arithmetic, one in computation and one in reasoning. The last test is in dictation, which is really a spelling test. In the advanced examination there are also three tests in reading, two in arithmetic, another in nature study and science, another in history and litera-

ture, another in language usage and the last is in dictation or spelling.

The primary test requires a total working time of about 60 minutes, which may well be broken up into two periods. The advanced examination requires a total working time of about 125 minutes and may be broken up into two or three periods. The time required to give the complete examination is believed by the authors to approximate the minimum which is consistent with the requisite degree of reliability. The total expenditure of time for assembling, giving and scoring the tests, and for the treatment of results is only a fraction of that which hitherto has been necessary "and much less than the time generally given to final examinations." Nevertheless, the time allowances are in all cases liberal enough to make the test almost entirely a test of *power* rather than *speed.* As a measure of the reliability of the tests the two forms of the test (Forms A and B) have been correlated with each other. The correlations ranged from .75 to .96 and the median is above .90. The probable error of the complete examination is approximately two months, that is, the chances are even that any age (e. g., 12 yrs., 6 mos.) as obtained by the Stanford Test is not in error by more than two months and the chances are twenty to one that it is not in error by more than 6 months.

The authors state [3] that "if the pupils in a school were classified upon the results of the Stanford Achievement Test alone, probably not more than 3 per cent of the 12 year old pupils would be placed in the wrong grade. This statement is all the more striking when we consider that in the four California school systems upon which our

[3] Manual, pp. 16-17.

TEST 5. ARITHMETIC: REASONING

Find all the answers as quickly as you can.

Write the answers on the dotted lines.

Use the blank sheets of paper to figure on.

Begin here.

1 How many are 5 birds and 4 birds? *Answer*.....

2 Three apples and two apples are how many apples? *Answer*.....

3 Jane bought a ruler for 5 cents and a bottle of ink for 8 cents. How much did she spend for both? *Answer*.....

4 How many days are there in 2 weeks? *Answer*.....

5 Mary had eight oranges and ate two. How many did she have left? *Answer*.....

TEST 6. NATURE STUDY AND SCIENCE

Samples: The number of cents in a dollar is 200 <u>100</u> 300

Our rain comes from the <u>clouds</u> moon stars

Draw a line under the word that makes the sentence true.

Begin here.

1 Christmas comes in December January July..... 1

2 The month before April is March May June..... 2

3 A calf is the young of the cow - goat horse..... 3

4 Soap is made from fats lemons sugars..... 4

5 Horseshoes are made of copper lead iron..... 5

TEST 7. HISTORY AND LITERATURE

Draw a line under the word that makes the sentence true.

---

1 The man who slept for 20 years was Ichabod Crane Miles Standish Rip Van Winkle.. 1

2 America was discovered by Balboa Columbus Hudson..... 2

3 Black Beauty was a crow dog horse..... 3

4 A famous American poet was Cooper Longfellow Shelley..... 4

5 The girl who ran down a rabbit hole was Alice Isabel Rosamund..... 5

TEST 8. LANGUAGE USAGE

Samples

Apples $\frac{\text{is}}{\underline{\text{are}}}$ good.

He $\frac{\underline{\text{told}}}{\text{telled}}$ me.

1 She was just about to $\frac{\text{sit}}{\text{set}}$ down.

2 I will $\frac{\text{teach}}{\text{learn}}$ him to do better.

3 There was a large $\frac{\text{mob}}{\text{crowd}}$ at church.

4 Four men and a boy $\frac{\text{are}}{\text{is}}$ in the party.

5 Jane is $\frac{\text{more prettier}}{\text{prettier}}$ than Helen.

FIG. 16. STANFORD ACHIEVEMENT TEST. ADVANCED EXAMINATION —FORM B. THE FIRST FIVE EXERCISES IN EACH OF THE FIRST EIGHT TESTS

TEST 1. READING: PARAGRAPH MEANING

Sample: Dick and Tom were playing ball in the field. Dick was throwing the ball and ..... was trying to catch it.

Write JUST ONE WORD on each dotted line.

1 Jack got his hat and ran to the door. "Where are you going?" said his mother. "To school," said ....., and ran off as fast as he could go.

2 Bess has a dog and a kitten, but her two pets do not like each other very well. When the dog comes near, the ..... always runs away as fast as it can.

3 Ned was crying because his little pony had died. Just then a fairy appeared and asked him why he was so sad. "Because," said Ned, "my dear little ..... is dead."

4 One day a lazy owl came to the magpie and begged her to build a nice nest for her. "Why should I build you a nest?" said the magpie. "If you were not so ..... you would build it yourself."

5 A gray pussy saw a lark out in the field and thought it would make a fine dinner. "Come here, pretty lark," said the ....., "and I will show you the bell that hangs on my neck." But the wise lark said he did not care to see the ..... and flew quickly away.

TEST 2. READING: SENTENCE MEANING

Samples: Can dogs bark?..... Yes No
Does a cat have six legs?..... Yes No

Read each question and draw a line under the right answer.

1 Do birds sing?	Yes	No	1
2 Do boys eat bread?.....	Yes	No	2
3 Do people have three feet?	Yes	No	3
4 Can a horse run a mile?.....	Yes	No	4
5 Do little girls ever laugh?.....	Yes	No	5

TEST 3. READING: WORD MEANING

Samples: Bread is something to catch drink eat throw wear
A robin is a bird cat dog girl horse

In each sentence draw a line under the word that makes the sentence true.

1 A teacher is a	boy family person school table...	1
2 Tears come when we	cry drink eat talk walk...	2
3 A tail is part of a	book cat face mountain week..	3
4 An oak is a kind of	box corn egg money tree....	4
5 A wheel is part of an	arm river train wall word.	5

TEST 4. ARITHMETIC: COMPUTATION

Get the answers to these examples as quickly as you can without making mistakes.
Look carefully at each example to see what you are to do.

Begin here.

(1)	(2)	(3)	(4)	(5)
		Add	Add	Add
2 + 4 =	5 + 5	5	15	4
		1	2	3

FIG. 17

norms are based, approximately 65 per cent of the 12 year old pupils are at present in the wrong grade, in the sense that in educational attainment they are nearer to the average attainment of a higher or lower grade than to the average of the grade in which they are actually found."

Method of Giving and Scoring the Tests—Complete details for giving the tests are contained in the Manual which accompanies the tests. There are no special difficulties in giving the tests except for the dictation tests and here it is only necessary to exercise care in slow distinct enunciation of the phrases read to the pupils. The Manual also contains general rules for scoring the tests. Special keys are provided for scoring each of the separate tests. The tests are strictly objective and with these keys the scoring becomes a matter of routine.

The score for each subject is first obtained in points. These point scores are to be transposed into educational scores called Subject Ages. For example, if a pupil's total score in reading were 141, his subject age for reading would be 12 yrs. 3 mos. Tables are provided for transferring all point scores into subject ages. The pupil's *"Educational age"* is the age equivalent of the sum of all the subject scores. Tables are provided for transforming these composite point scores into educational age equivalents.

Function of the Tests—The Stanford Achievement Tests are by far the most satisfactory measurement of school achievement. Individual tests of the different school subjects have been described and their respective advantages pointed out. Each of these tests has its particular advantages and uses, but for a satisfactory general measure of school attainment the Stanford Tests

have combined most of the important features of several others and made uniform the nature and procedure of the tests so that the administrator or teacher may rank her pupils on the basis of attainment. As has already been pointed out in earlier chapters, there are decided advantages in standardized tests. In the Stanford Tests these advantages may be made most valuable in classifying pupils for instruction.

May we look forward to the time when such standard tests may be elaborated into enough forms and extended so as to include the junior and senior high school subjects. Then promotions may become standardized, and will be less a matter of personal judgment or even of personal prejudice, as is now too often the case.

Materials Needed

Kelley, T. L., Ruch, G. M., and Terman, L. M., The Stanford Achievement Test for grades 2-8. Primary Examination for grades 2 and 3, Advanced Examination for grades 4-8. World Book Co., Yonkers, N. Y.

Specimen Set. An envelope containing 1 copy of each examination (both forms), each Key, a Manual, and a Class Record. Price 50 cents postpaid.

Selected References

Stanford Achievement Test by Truman L. Kelley, Giles M. Ruch and Lewis M. Terman, Manual of Directions, World Book Co.

Monroe, W. S., The Illinois Examination, Bulletin No. 6, Bureau of Educational Research, University of Illinois, Urbana, Ill.

Pintner, R., and Marshall H., "A Combined Mental-Educational Survey," *Journal of Educational Psychology*, Jan. & Feb. 1921.

CHAPTER XVII

INTELLIGENCE TESTS

In Chapter 3, a statement was made regarding the uses of intelligence tests in the classroom, and it was there pointed out that the teacher can make good use of the tests as diagnostic instruments. To be sure, many psychologists are still of the opinion that a more extended knowledge of psychology is needed than most classroom teachers possess, in order to interpret the results of such tests in a way to avoid misuse. But if directions are followed carefully, and care is taken not to give too much importance to the result of any single test, without corroborative evidence, the use of a selection of tests from the list to be discussed in this chapter will give valuable information.

Teachers need not be too much disturbed over the controversies and discussions which have filled educational and lay periodicals alike for the past two years as to the use of the tests and as to what they actually do measure. Thus, some authors do not admit the term "intelligence" in connection with the tests, but say they are measures of "mentality," thus differentiating sharply between "intelligence," as something which may be acquired or developed, and "mentality," which is something native and may not be developed beyond a certain fixed point. Others do not make this distinction, but use the two

expressions interchangeably. In fact, no one has yet been able to give a definition of "intelligence" which is universally acceptable, and so there is now a great deal of confusion in the minds of people generally as to what the "intelligence" tests really do measure Perhaps the best statement that has been made is that they measure to a really marked degree, *ability to do school work.* The classroom teacher will certainly be satisfied with this statement, in so far as the actual function of the test is concerned, and so will not be more disturbed over the true nature of "intelligence," than over the as yet undetermined answer to the question "What is electricity?"

Two sorts of tests are in general use, ordinarily known as individual and group tests. The teacher of limited psychological training will not attempt very much with the individual tests; the group test she may use to some advantage. A brief discussion of both kinds is, however, in order.

Individual tests are those which are given to a single pupil at a time, under conditions which isolate him from other pupils for the period of the test, and which make possible the elimination of any distracting factors. These tests as now given in this country are adaptations of the series of tests devised by Messrs. Binet and Simon in France during the years 1905 to 1911.

American revisions of the Binet scale have been made by Messrs. Goddard, Yerkes, Terman, and others. Of these, the Terman revision, known generally as the Stanford Revision of the Binet Scale, is the most widely used. This "revision" consists of ninety tests, arranged to be given to children from the age level of three years up to that of the "superior adult." They are fully described

in Terman's volume entitled *The Measurement of Intelligence* and their applications are explained in his further volume *The Intelligence of School Children.*

These volumes are to be accompanied by test materials and instruction booklets, and require careful study and guidance by an experienced teacher before they can be used with any guarantee of accuracy. Under proper conditions of application and interpretation, they are conceded to be the best measure yet devised for general learning ability of children of elementary and intermediate school age. For children of high school age, and adults, opinion is divided as to their being a better measure than some other tests. In general, the high school teacher may feel that they do give as accurate a test for most pupils, as any other measure, but that this statement will not hold for all cases. And the caution must be made here, as has already been indicated in this volume, that no worse mistake can be made than to attempt any determination of a child's school grading, selection of studies, or other relations with the school or with life outside of the school on the single basis of the intelligence test. Heredity, biological and social environment, previous training, disposition, health, physical characteristics, moral and social qualities, are all to be taken into consideration before a final conclusion is drawn.

Group tests are those which are arranged to be given to the entire class at one time, all pupils working simultaneously. The usual form is that of a booklet which contains a set of tests of varying character, each of which is to be completed separately, so far as can be done within certain time limits set by the author of the test.

Such tests were worked out originally with the idea of

presenting an adaptation of the Binet idea which would have the added advantage of application to large numbers of pupils within a relatively brief time. To give the Stanford Revision of the Binet tests to a single pupil takes from an hour to an hour and a half, in most cases. By shortening the procedure somewhat, it can be done in forty-five minutes, but such procedure may be open to inaccuracies. Thus approximately thirty hours would be required to give these tests to a class of thirty pupils. But these same thirty pupils may be given any of the group tests in from thirty to sixty minutes, and the results are sufficiently accurate to justify almost as satisfactory diagnosis as would be the case if the individual test were given. So the individual tests are ordinarily reserved to be given in special cases presenting particular difficulties, and the usual mental measure is the group test.

The first of these tests to have widespread use was that of Dr. Arthur S. Otis, who prepared a group of tests to be used as a rough measure of intelligence, in 1917. His idea was made the basis for the form, and to an extent the content, of the tests worked out by the group of psychologists who made up the tests used by the United States Army during the war. After the war, Dr. Otis, profiting by the experience of the war, revised his tests, and they are now considered one of the best group measures. They consist of two parts, the Primary Examination and the Advanced Examination. The first is designed for grades 1 to 4, and the second for grades 5 to 12. The Primary is made up of eight tests, of which six are nonverbal, and takes 25 minutes to apply. The Advanced consists of ten tests, as follows: 1, Following Directions; 2, Opposites; 3, Disarranged Sentences; 4, Proverbs; 5,

Arithmetic; 6, Geometric Figures; 7, Analogies; 8, Similarities; 9, Narrative Completion; 10, Memory. The latter takes 60 minutes to apply. Each of the tests has two forms, so that a second form may be given if the first seems not to give satisfactory results.

The tests used during the war were the cause of a tremendous interest in intelligence testing, and were the means of a widespread utilization of similar tests in industry as well as in education. The best known form of the army tests was called the Army Alpha, and although constructed for military purposes, it has proved a very fair measure of general ability in the schools and colleges in which it has been given. However, it is not recommended for school use, especially below the high school, as long as there are many excellent measures which have been constructed for special school testing.

In addition to the Otis test, other tests which are widely used in public schools are: The National Intelligence Tests, prepared by a group of the best known psychologists in America, Messrs. M. E. Haggerty, L. M. Terman, E. L. Thorndike, G. M. Whipple, and R. M. Yerkes. These consist of two scales of five tests each, Scale A involving Arithmetic Problems, Sentence Completion, Logical Selection, Synonym-Antonym, and Symbol-Digit tests, and Scale B, Computation, Information, Vocabulary, Analogies, and Comparison. Several forms of each scale are available. The test is for use in grades 3 to 8, inclusive, and takes from 30 to 35 minutes to give.

The Dearborn Group Tests of Intelligence consist of two series, one for grades 1 to 3, and the other for grades 4 to 9. The first consists of a pictorial series, non-verbal in nature, and requires three periods of 25 minutes each

for giving. The second consists of ten tests: 1, Picture Sequences; 2, Word Sequences; 3, Form Completion; 4, Opposite Completion; 5, Memory Ladder; 6, Picture Symbols; 7, Mazes; 8, Disarranged Proverbs; 9, Faulty Pictures, and 10, Number Problems. There are two forms of these, each taking 50 minutes for giving. As seven of the ten tests are non-verbal, the test as a whole differs in form very markedly from the National and the Otis, which are largely verbal.

The Pintner-Cunningham Primary Mental Test is another more recent test for the kindergarten, first and second grades. There are seven parts to the test, all of which are composed of pictures. First, the pupil marks certain parts of the pictures as directed by the oral instructions of the examiner; second, he picks out the prettiest picture from two sets of similar pictures; the third part is an Associated Objects test; the fourth is a Discrimination of size test; the fifth is a Picture Parts test; the sixth is a Picture Completion test, and the seventh is a Dot Drawing test. Pintner found that this test correlated very high with both the teacher's ranking of her pupils (+.64 to +.78) and the scores of the Stanford Revision of the Binet-Simon scale (+.55 to +.82).

The Haggerty Intelligence Examination, Delta 1, is used in grades 1 to 3, and Delta 2, in grades 4 to 9. Delta 1 consists of six tests, of which five are non-verbal. It requires 30 minutes for giving. Delta 2 requires 35 minutes, and is made up of the following six tests; Sentence Reading, Arithmetic, Picture Completion, Synonym-Antonym, Common-sense, General Information.

The Mentimeters, by Trabue and Stockbridge, are applicable to all persons from infants to the university, the

time for giving them varying with the group to which they are applied. Typical Mentimeters comprise: Pictorial Absurdities, Maze Threading, Geometric Figures, Opposites, Reading Directions, Completion, Arithmetic, Range of Information.

The Illinois Examination, by Monroe and Buckingham, Form I, for grades 3, 4, 5, and Form II, grades 6, 7, 8, takes in each case about 60 minutes to give. This involves not only intelligence tests, but also the giving of arithmetic tests and the Monroe Reading Test. The intelligence tests comprise: Analogies, Arithmetic Problems, Sentence Vocabulary, Substitution, Verbal Ingenuity, Arithmetical Ingenuity, Synonym-Antonym.

The Terman Group Test of Mental Ability is designed for grades 7 to 12. It consists of ten tests: Information, Best Answer, Word Meaning, Logical Selection, Arithmetic, Sentence Meaning, Analogies, Mixed, Classification, Number Series. It takes 35 minutes in application.

The Myers' Mental Measure is designed for all grades into the university. It takes but 20 minutes for application, as it consists of four pictorial tests, involving no verbal tests.

This is by no means a complete list of the mental measures which are in use; but it gives a sufficient range and variety to enable the teacher to use several different types of test on any class, and so to get a kind of data which would be valuable in a complete diagnosis. Directions for giving accompany each of these tests, together with keys for scoring, and directions for evaluating the results. The scoring is a very important feature in giving the tests, and should not be attempted without complete directions. Usually aids, such as celluloid sheets

or oiled paper or cardboard guides, are used to score, involving the stencil idea, and thus the work of correction is reduced to a very slight task. The norms or other means of evaluating the work of the group, or of the individual within the group, are also of marked importance, and the value of the tests depends upon the evaluation of the scores in the light of these norms.

The reader is referred for the best symposium yet published on the Intelligence Test, to the Twenty-First Yearbook of the National Society for the Study of Education, published in 1922 by the Public School Publishing Co., of Bloomington, Ill. Part I deals with the nature, history, and general principles of intelligence testing, and Part II with the administrative use of intelligence tests.

As a final word, the teacher is advised to use these group tests only when they are not given to her pupils through any other agency; but if she has no means of getting reliable information about her class, the giving of the tests will enable her to make at least a partial estimate of the general ability of the class to do school work, and will enable her to determine whether she is getting the results from the pupils which she is justified in expecting in line with their ability as indicated by the mental measure.

Materials Needed

Dearborn, W. F., Group Tests of Intelligence, Series I, grades 1-3. Series II, grades 4-9. J. B. Lippincott and Co., Philadelphia, Pa. Specimen Set 15 cents, either series 25 booklets $1.50. Examiners' Guide 10 cents.

Haggerty, M. E., Intelligence Examination, Delta 1 grades 1-3, Delta 2 grades 3-9. World Book Co., Yonkers, N. Y. Speci-

men Set 55 cents, Delta 1 $1.30 for 25 booklets, Delta 2 $1.25 for 25 booklets, Manual of directions 25 cents.

Haggerty, M. E., and others, The National Intelligence Tests. World Book Co., Yonkers, N. Y. Scale A, form 1, Scale B, form 2, are all alternative forms. Any form complete 1.30 for 25, Manual of Directions, 20 cents.

Monroe and Buckingham, The Illinois General Intelligence Scale, Illinois Examination I for grades 3-5, Illinois Examination II for grades 6-8, forms 1 and 2 of each. Public School Publishing Co., Bloomington, Ill. Sample Set 20 cents, $2.00 per hundred.

Myers, G. C., Mental Measure, for all grades. Newson and Co., 73 Fifth Ave., New York City. Sample Copy 10 cents, 12 copies $1.00, 100 copies $5.00. Manual 80 cents.

Otis, A. S., Group Intelligence Scale, Primary Examination for grades 1-4, Advanced Examination for grades 5-12, forms A and B of each. World Book Co., Yonkers, N. Y. Specimen Set 50 cents. Primary examination $1.25 for 25, Advanced Examination $1.30 for 25. Manual 30 cents.

Pintner and Cunningham, Primary Mental Test for kindergarten to second grade. World Book Co., Yonkers, N. Y. Specimen Set 20 cents, $1.45 for 25 including Manual.

Terman, L. M., The Measurement of Intelligence (individual examination) for any age above 3 years. Houghton Mifflin Co., New York City. Condensed Guide $1.00, Test material $1.00, 25 Record Booklets $2.00, Abbreviated Filing Cards $1.00 per hundred.

Terman, L. M., Group Test of Mental Ability for grades 7-12, forms A and B. World Book Co., Yonkers, N. Y. Specimen Set 15 cents. Either form complete $1.35 for 25.

Trabue and Stockbridge, Mentimeters for any grade. Doubleday, Page and Co., Garden City, L. I. Specimen Set 25 cents, $1.75 for 25 pupils complete.

Selected References

Terman, L. M., *The Measurement of Intelligence,* Houghton Mifflin Co.

Terman, L. M., *The Intelligence of School Children,* Houghton Mifflin Co.

Pintner, Rudolf, *Intelligence Testing,* Henry Holt and Co.

Twenty-First Yearbook of the National Society for the Study of Education, Public School Publishing Co., Bloomington, Ill.

CHAPTER XVIII

STATISTICAL AND GRAPHIC METHODS

Several terms have been used in discussing the results of measurements in the preceding chapters that may not have been clear to those unfamiliar with statistics and statistical methods. It is the purpose of this chapter to present the simpler facts of statistics and of graphic methods in such untechnical language that those unfamiliar with the principles may obtain an understanding of them.

The final value of any program of measuring the results of teaching depends upon the information gained as the result of the measurements. Such information can be gained only from an orderly presentation of the results obtained. Too often tests have been given with a feeling that there is a virtue in the tests themselves and the results are too often filed away without any further use being made of them. This has generally occurred because the teacher or supervisor did not know how to deal with and interpret the data. There are two ways of dealing with data, one of which is called the statistical and the other the graphic method. Statistics mean the manipulation of the numerical values obtained by measurement in such a way as to present the results briefly and intelligently. These same data may also ordinarily be

presented by means of drawings or graphs in such a way as to show these same facts and results in another form.

Complicated as statistical methods may be, the main purpose is generally to show one or more of three general facts about series of measures. The first of these is the general, or what is technically called the central, tendency of the measures. This central tendency may be considered as the typical case within the group. It is the single value which most nearly represents the group. In the tests it is called the "Norm."

Measures of central tendency—There are three different measures of central tendency. These are called the average, the median and the mode. The average may be defined as "the sum of the values of all the measures in the distribution, divided by the number of measures."[1] This term and the method of obtaining it is familiar to every teacher and needs no further explanation. The median is the *middle* measure in a series and is found by arranging the scores of a series in order of magnitude. The middle value is the median. If there are 35 scores in a series, the eighteenth value is the median. The mode is the most frequent value in a series.

If a large enough number of cases is included in any series of measurements, the scores tend to distribute themselves around some central tendency so long as the measure represents a random sampling. For example, if all the men in a college were arranged in order of height, the tallest would probably be over 6 feet and the shortest probably would not be 5 feet. The average would be about 5 ft. 10½ inches. This is one measure

[1] H. O. Rugg, *Statistical Methods Applied to Education*, p. 115, Houghton Mifflin Co.

of the central tendency. If the group is unselected, that is, if the group is representative of the adult male population in general, the height of the middle man in the company will also be about 5 ft. 10½ inches. There are also likely to be more men of this height than any other. Whenever we have a large enough number of cases and the cases are unselected, the average, the median and the mode will be approximately the same value. In such a case it will not make much difference which measure of central tendency is used.

These terms may be made clearer from the following illustration. Suppose the grades in a class of 21 pupils in arithmetic are:

A....82	E....83	I....90	M....98	Q....79	V....81
B....79	F....79	J....80	N....48	R....60	
C....35	G....75	K....79	O....75	S....95	
D....96	H....51	L....82	P....78	T....50	

The average score for the class is the sum of the scores, 1575, divided by the number of scores, 21, which is 75. The median is the middle score. To find the median arrange the scores in order of magnitude. 35, 48, 50, 51, 60, 75, 75, 78, 79, 79, 79, 79, 80, 81, 82, 82, 83, 90, 95, 96, 98. The median is the eleventh case, which is 79. The mode is the most frequent score in the given series. It is also 79, since there are four scores of this value and only two scores of any other single value.

The methods described thus far are satisfactory for finding the central tendency when the number of cases is small. When there is a large number of cases it is advisable to condense the data into a table of frequency. For example, if there are four scores of 79 it is not necessary to write 79 four times, but arrange the results in tabular

STATEMENT AND METHOD OF SOLUTION OF PROBLEM I

Pupils	*Raw Scores*	*Scores Arranged in Order of Magnitude*
A	82	35
B	79	48
C	35	50
D	96	51
E	83	60
F	79	75
G	75	75
H	40	78
I	90	Mode (the most frequent score) 79
J	80	79
K	79	79 ← Median
L	82	7((the middle score)
M	98	80
N	48	81
O	75	82
P	78	82
Q	79	83
R	60	90
S	95	95
T	50	96
U	81	98
Sum	1575	

$$\frac{1575}{21} = 75 \text{ Av.}$$

Number of Cases 21

Steps in the Calculation of the Average When the Data Are Not Arranged in a Table of Frequency.

$$\text{Av.} = \frac{\text{Sum of scores}}{\text{No. of cases}}$$

Find the sum of all the scores and divide this sum by the number of scores.

Steps in the Calculation of the Median When the Data Are Not Arranged in a Table of Frequency.

$$\text{M.} = \frac{(\text{No. of cases} + 1)\text{th Measure}}{2}$$

Arrange the scores in order of magnitude. The median case is found by counting one-half case more than half the number of cases beginning at either the highest or lowest score.

form under the headings "score" and "number of cases." The following is an illustration of a table of frequency.

PROBLEM II

Scores of a Class in History	*Number of Cases*
32	1
33	2
38	1
41	1
55	2
64	1
72	3
77	4
78	6
82	8
86	2
90	3
92	2
94	1
96	1
Number of Cases	38

Steps in the Computation of the Average When the Data Are Arranged in a Frequency Table with 1 as the Unit of Frequency

$$\text{Av.} = \frac{\text{Sum of (Score} \times \text{Frequency)}}{\text{Number of Cases}}$$

Multiply each score by its frequency (number of times it occurs) and divide the sum of the products by the number of cases.

Steps in the Computation of the Median When the Data Are Arranged in a Table of Frequency with 1 as the Unit of Frequency

M = The middle measure.

First: Compute N/2 measures, that is, divide the number of measures by 2.

Second: Beginning at either the highest or lowest score count the

number of measures included in the column of frequencies to the frequency value that contains the median.

Third: From N/2 measures subtract the total number below the frequency value as obtained in step 2. This number of measures is the number that is needed to be included from the next frequency to bring the computation to the median point on the scale.

Fourth: Divide this remainder by the number of frequencies in the interval (that is, the frequency value within which the median lies). This is the proportion of the total measures in the interval that is needed to bring the computation to the median point.

Fifth: Multiply this ratio by 1 (the unit of frequency). The product is then to be added to the score below the middle measure if the calculation is from the lowest score or subtracted from the one just above the middle measure if the calculation is from the highest measure.[1]

The average of the scores in such a frequency table is found by multiplying each score by its frequency and dividing the product by the number of cases. $(32 \times 1) + (33 \times 2) + (38 \times 1) \text{ ---------- } + (96 \times 1) = 2815$. $2815 \div 38$, the number of cases, $= 74$ 3-38, the average.

The median is the middle measure. In this problem it lies somewhere between 77 and 78. It may be found by the following method. Beginning at the lower end of the table add the frequencies to the value equal to or just below one-half the number of cases. That is, do not add the frequency which makes the total more than half the cases. If the sum of the frequencies is just equal to one-half the number of cases, the highest score of the last frequency included in the addition is the median. In case the sum of the frequencies is less than one-half the

[1] Adapted from H. O. Rugg, *Statistical Methods Applied to Education*, p. 113.

number of cases, the median is derived by interpolation as illustrated from the data of Problem II. $1 + 2 + 1 + 1 + 2 + 1 + 4 + 4 = 15$. The median is four cases farther, that is, it includes 4/6 of the next group of cases, or is $4/6 \times 1$ (the unit of frequency) $= 2/3$ of the way within the next score. The median is, therefore, $78 + 2/3 = 78\ 2/3$. The mode is the most frequent score, which in this problem is 82.

Sometimes it is desirable to condense the data even more than in the method already described. Instead of giving the actual scores, they may be grouped into class intervals. That is, instead of making the unit of frequency 1, it may be 2, 5, 10 or any other number. In the foregoing Problem II, if the scores were grouped with 5 as the unit of frequency, the results would be stated as follows:

PROBLEM III

Class Interval	*Midpoint*	*Frequency*
30-34	32	3
35-39	37	1
40-44	42	1
45-49	47	0
50-54	52	0
55-59	57	2
60-64	62	1
65-69	67	0
70-74	72	3
75-79	77	10
80-84	82	8
85-89	97	2
90-94	92	6
95-99	97	1

The average is found in much the same way with this arrangement as in the preceding problem. Here the mid-

point of each class interval is multiplied by the frequency and the sum of the products divided by the number of cases $(32 \times 3) + (37 \times 1) \text{ ---------- } + (97 \times 1)$ divided by $38 = 74$.

Method of Determination of the Average and Median for Problem III

Class Interval	*Midpoint*	*Frequency*	*Product*
30-34	32	3	96
35-39	37	1	37
40-44	42	1	42
45-49	47	0	0
50-54	52	0	0
55-59	57	2	114
60-64	62	1	62
65-69	67	0	0
70-74	72	3	216
75-79	77	10	770
80-84	82	8	656
85-89	87	2	174
90-94	92	6	552
95-100	97	1	97
		Sum	2816

$\frac{2816}{38} = 74.1$ Av.

Number of Cases 38

Steps in the Computation of the Average When the Data Are Arranged in a Frequency Table with Class Intervals

First: Determine the midpoint of each class interval.

Second: Multiply these midpoints by their frequencies.

Third: Divide the sums of the products of step 2 by the number of cases. The quotient is the average.

Steps in the Computation of the Median When the Data Are Arranged in a Frequency Table with Class Intervals

The first four steps are the same as in the calculation of the median for Problem II.

Fifth: Multiply this ratio by the number of units in the class interval. The product is the number of units on the scale that need to be added to the value of the lower limit of the class interval to give the median.

Sixth: Add this number to the value of the lower limit of the class interval. This is the median point on the scale.

In general, when the group is not large and not representative, the mode is the least reliable of the three measures. If we are interested in the individual case most representative of the group, the median is the better measure. It is least influenced by extreme variation. For example, in Problem I the 21 class grades in arithmetic, the median would be the same, 79, if the lowest score were 14 instead of 35. On the other hand, the average is affected by the size of every score in the series and a reduction of the first score from 35 to 14 would make the average 74 instead of 75. The question as to which of the measures of central tendencies should be used depends upon what is to be shown by the data. Sometimes one measure and sometimes the other should be used.

Measures of Deviation

Thus far we have dealt only with the measure of the central tendency or typical measure of a series. But data cannot be explained by a simple and single measure such as the central tendency. A further problem arises as to how the data arrange around this typical measure. Are the values closely arranged or widely distributed and how much? The measure of the variation from the central tendency is called the deviation, dispersion or variability. There are three common measures of deviation. They are the average deviation, the median deviation or

probable error, and the standard deviation. Any or all of these values may be calculated from any one of the three measures of central tendency. In the data of Problem I the average deviation from the average may be calculated by determining how much each measure deviates from the average and by dividing the sum of the deviations by the number of cases. Although it is not essential for such calculations, the data should be arranged in order of magnitude.

CALCULATED DEVIATIONS FOR PROBLEM I

The Average is 75

Arithmetic Scores	*Deviation*	*Arithmetic Scores*	*Deviation*
35	40	79	4
48	27	80	5
50	25	81	6
51	24	82	7
60	15	82	7
75	0	83	8
75	0	90	15
78	3	95	20
79	4	96	21
79	4	98	23
79	4		

The average deviation is the sum of $40+27+25\ldots\ldots+21+23$ divided by 21 which equals 12 10/21.

The median deviation[1] is the middle measure of the deviations. The deviations arranged in order are 0, 0, 3, 4, 4, 4, 4, 5, 6, 7, 7,

[1] The probable error which is one half the difference between the ¼ measure and the ¾ measure—the median being the ½ measure—is approximately the same as and will equal the median deviation when the distribution is symmetrical. It also includes approximately one half the cases. The formula is $Q=\frac{Q_1-Q_3}{2}$ in which Q is the quartile deviation or probable error, Q_1 the ¼ quartile and Q_3 the ¾ quartile.

8, 15, 15, 20, 21, 23, 24, 25, 27, and 40 and the median deviation is therefore, 7. This deviation on either side of the average includes approximately one-half the total number of measures. The standard deviation is the square root of the sum of the squares of all the deviations over the number of cases.

$$\frac{\sqrt{40^2+27^2+25^2+\ldots\ldots\ldots\ldots 21^2+23^2}}{21} = 16.4.$$

The different measures of variability may also be determined from data arranged into a table of frequency. In Problem III the average deviation is found by dividing the sums of the deviations of the midpoints of each class interval times the frequency by the number of cases. In this case we will take, for example, the deviations from the median, which is 79. The deviations multiplied by the frequencies for Problem III are as follows:

Method for the Determination of the Average Deviation for Problem III

Class Interval	*Midpoint*	*Median*	*Deviation*	*Frequency*	*Product*
30-34	32	79	47	× 3	= 141
35-39	37	79	42	× 1	= 42
40-44	42	79	37	× 1	= 37
45-49	47	79	32	× 0	= 0
50-54	52	79	27	× 0	= 0
55-59	57	79	22	× 2	= 44
60-64	62	79	17	× 1	= 17
65-69	67	79	12	× 0	= 0
70-74	72	79	7	× 3	= 21
75-79	77	79	2	× 10	= 20
80-84	82	79	3	× 8	= 24
85-89	87	79	8	× 2	= 16

Method for the Determination of the Average Deviation for Problem III—Continued

Class Interval	*Midpoint*	*Median*	*Deviation*	*Frequency*		*Product*
90-94	92	79	13	× 6	=	78
95-100	97	79	18	× 1	=	18
				Sum		458

$$\frac{458}{38} = 12.05 \text{ A. D.}$$

Number of Cases 38

Steps in the Computation of the Average Deviation (from the Median) When the Data Are Arranged in Class Intervals

First: Determine the midpoint of each class interval.

Second: Subtract each midpoint value from the median (or average as desired).

Third: Multiply each deviation by its frequency.

Fourth: Divide the sums of the products obtained in 3 by the number of cases. This gives the average deviation.

The probable error in this problem is calculated by dividing the difference between the first and third quartiles by two. The first quartile is one-fourth the way down within the series and the third quartile is three-fourths of the way down. The first quartile, therefore, is the 9½ case and the third quartile is the 28½ case. The position of the quartiles is found in the same way as the median, which may be considered the second quartile. The first quartile is

$$70 + \frac{9\frac{1}{2} - (3+1+1+0+0+2+1+0)}{3} \times 5 = 72.5$$

The third quartile is

$$80+\frac{28\frac{1}{2}-(3+1+1+0+0+2+1+0+3+10)}{8}\times 5=84.7.$$

$\frac{84.7-72.5}{2}=6.1$ the P. E. or, as it is sometimes called, the quartile deviation.

The standard deviation is computed by the formula $\sigma=\frac{\sqrt{\Sigma fd^2}}{N}$ σ = standard deviation, Σfd^2 = the sum of the deviations squared and multiplied by the frequencies, N = the number of cases. In obtaining the S. D. in Problem III we have

Deviation	d^2	f	fd^2	D	d^2	f	fd^2
47	2209	3	6627	12	144	0	0
42	1764	1	1764	7	49	3	147
37	1369	1	1369	2	4	10	40
32	1024	0	0	3	9	8	72
27	729	0	0	8	64	2	128
22	484	2	968	13	169	6	1014
17	289	1	289	18	324	1	324
							12742

$$\frac{\sqrt{12742}}{38}=18.31.$$

Correlation

There is one other important measure that must be used in comparing two series of data. The central tendency and the deviation may describe a single series of classroom scores but it gives no adequate method of relating scores in one test with scores in another subject. A class is given a test in history and another in

arithmetic. How do the scores in one subject compare with the scores in the other? We may determine the central tendency and the deviation for each set of scores, but this will not tell whether those who are high in one subject are high or low in the other. A simple method of such comparison consists in dividing each of the two series of scores to be compared into deciles (10 percentiles) and comparing decile standings. For example, if we find that most of the children in the highest decile in history are also in the highest decile in arithmetic and that those in the second decile in one are mostly in the second decile of the other, et cetera, we conclude that there is a close relationship in the two sets of scores. On the other hand, if those in the highest percentiles in history are distributed within several percentiles in arithmetic and this holds throughout, there is little relationship between the two sets of scores. If it is desired to make a quantitative measure of this relationship, another statistical procedure is necessary. The method of measuring this relationship is called correlation.

Correlation may be either positive or negative. Positive correlation means that the grades in the two series being compared vary directly, that is, when one is high the other tends to be high, when one is low the other tends to be low. Negative correlation means just the opposite. When one is high the other tends to be low, and vice versa. No correlation means no relationship. Complete positive correlation is expressed by $+1$ and complete negative correlation by -1.

There are two common methods of correlation named from their authors. We will consider only the Pearson

method. It is expressed by the formula $r = \frac{\Sigma x' y'}{N\sigma_x\sigma_y}$ in which r = correlation; $\Sigma x' y'$ = the sum of the products of the deviations of the two series; and $N\sigma_x \times \sigma_y$ = the products of the standard deviations of the two series multiplied by the number of cases. In practice $N\sigma_x \times \sigma_y$ may be simplified into $\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}$, that is, the square roots of the sums of the squares of the two series of deviations may be multiplied together for the denominator of the fraction. This eliminates dividing each sum of the squares by the number of cases and then multiplying again by the number of cases. The procedure may be made clear from the following computation.

PROBLEM IV

Pupil	*Arithmetic Grades*	*History Grades*	dx	dy	$d_x x$	$d_y y$	$x'y'$
C	40	55	—20	—15	400	225	+ 300
D ...	45	50	—15	—20	225	400	+ 300
r	48	30	—12	—40	144	1600	+ 480
g	50	50	—10	—20	100	400	+ 300
b	54	82	— 6	+12	36	144	— 72
j	60	70	0	0	0	0	0
a	72	94	+12	+24	144	576	288
h	84	45	+24	—25	576	625	— 600
e	90	86	+30	+16	900	256	+ 480
i	92	78	+32	+ 8	1024	64	+ 256
f	96	100	+36	+30	1296	900	+1080
Median	60	70		Sum	4845	5190	+2812

$$r = \frac{+2812}{\sqrt{4845}.\sqrt{5190}} = +.57.$$

Steps in the Computation of the Correlation (Pearson Method) of Two Sets of Scores for the Same Group of Pupils

First: Arrange the two sets of scores to be correlated in vertical columns, designated as x and y, with the two scores of each pupil opposite each other in the two columns.

Second: Find the average or median (ordinarily it makes little difference which is used) for each column.

Third: Compute the deviation for each column from the average or median for that column, observing signs. These are to be designated the dx and dy columns.

Fourth: Multiply each deviation of the dx column by the corresponding deviation of the dy column, observing signs. The algebraic sum of these products constitutes the Σd^2xy, usually written $\Sigma x'y'$ which is the numerator of the fraction for correlation.

Fifth: Square each of the values in column dx and column dy for the values for column d^2x and column d^2y.

Sixth: Find the sums of column d^2x and column d^2y, extract the square root of each sum and multiply one square root by the other. This gives the $\sqrt{\Sigma x^2}.\sqrt{\Sigma y^2}$ which is the denominator of the fraction for correlation.

Seventh: Divide the $\Sigma x'y'$ value obtained in step 4 by the $\sqrt{\Sigma x^2}.\sqrt{\Sigma y^2}$ value obtained in step 6 for the value of r.

The deviations from the average or median—median in the above example—are obtained as shown in the dx and dy columns. In this case plus and minus signs must be affixed to the deviations. These deviations are squared for the dx and dy columns. The signs in these columns are all plus. Deviation x is multiplied by deviation y, observing signs, for column xy. The algebraic sum of the xy column is the numerator for the value of r. The product of the square roots of the sum of the dx and dy columns is the denominator of the fraction.

As already stated there are other methods of determining correlation. A short method is provided when there

are large series of scores to be correlated. Some prefer the Spearman rank method of correlation for which the formula is:

$$\rho = 1 - \frac{6SD^2}{N(N^2-1)}$$

where D = any difference in the rank of an individual in the two series.

For a more complete description of these and other methods of correlation see some standard text in statistics, such as those already referred to in this chapter.

GRAPHIC METHOD

It is generally advisable not only to arrange the results of tests in statistical form but also to present them in graphic form. Often facts and relationships not apparent in the statistical material are brought out in the graphs. Graphic methods present few special difficulties and their widespread use makes it imperative that the class room teacher be familiar with the construction and reading of graphs.

One of the simplest forms of graphic method used in education and business is here presented to show the percentage of children having playgrounds of various sizes. This graph is self-explanatory.

Another common use of graphs is to show successive scores in some function. This is well illustrated by the well-known learning curve in telegraphy by Bryan and Harter. In this graph time is represented along the abscissa and the number of letters sent or received along the ordinate. The rate for any amount of practice is shown by where the score line cuts that ordinate. For example, the sending rate for the 12th week of practice

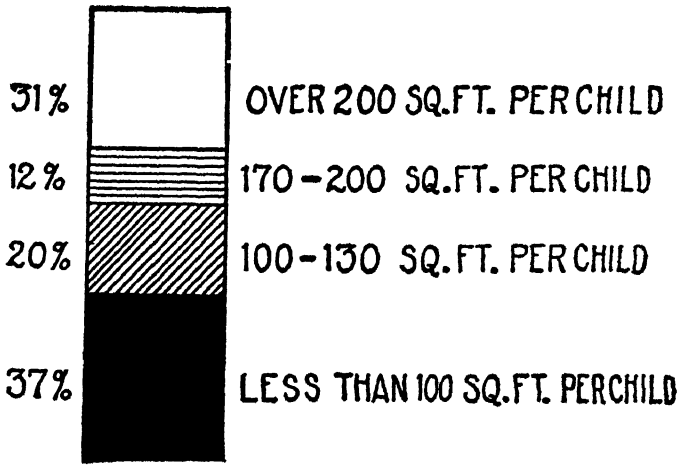


FIG. 18. SHOWING THE PERCENTAGE OF CHILDREN HAVING PLAY-GROUNDS OF VARIOUS SIZES *

* From Statistical Methods Applied to Education, H. O. Rugg, p. 359.

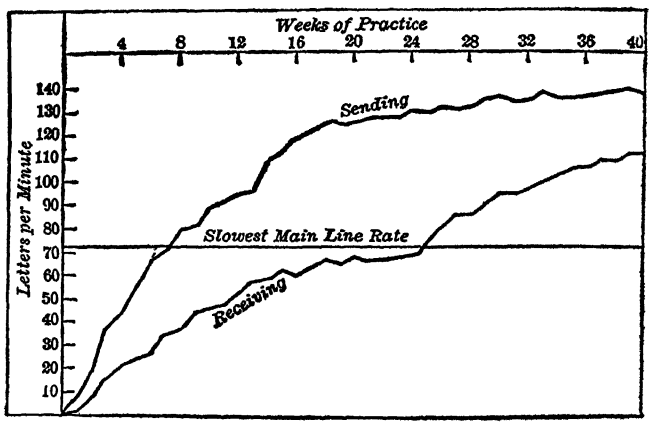


FIG. 19. CURVES OF LEARNING TO SEND AND RECEIVE TELEGRAPHIC MESSAGES (After Bryan and Harter)

is about 93 letters per minute and for the 24th week about 130 words per minute. In such curves custom has established the zero point at the lower left hand corner. Time or amount of practice is generally represented along the horizontal axis and units of work along the vertical axis.

The following data may well be constructed into graphic form to illustrate such construction. A pupil practiced silent reading 5 minutes per day for one month. His rate of reading in words read per second was as follows:

PROBLEM V

Day	Words per sec.	Day	Words per sec.	Day	Words per sec.	Day	Words per sec.
1	2.34	6	2.94	11	3.40	16	3.68
2	2	7	3.00	12	3.44	17	3.80
3	2.58	8	3.28	13	3.52	18	3.95
4	2.75	9	3.32	14	3.60	19	3.80
5	2.90	10	3.00	15	3.68	20	3.92

These data may be presented as shown in the graph on page 262.

These graphs show the usual characteristics of a learning curve. There is rapid progress at first, followed by less rapid gain during the latter part of the curve. At places in the graph there is practically no gain. These places are called "plateaus." Various reasons have been assigned for plateaus in learning. Swift [1] and others attribute them to waning of interest or attention. Bryan and Harter [2] explain plateaus as places where lower order habits are becoming automatized before more complex learning habits can be formed. After a certain

[1] Swift, E. J., *The Mind in the Making.*

[2] Bryan and Harter, "Studies in the Physiology and Psychology of the Telegraphic Language," *Psychological Review*, 4: 27-53 and 6: 348-375, 1897-9.

amount of practice a point is reached in which progress practically ceases. This is a permanent plateau and marks the physiological limit of improvement.

The results of a class may be shown as well as that of an individual. With the class the results are arranged from day to day or practice to practice. Unfortunately such graphs do not show deviations but only central tendencies. Several devices have been used in connection

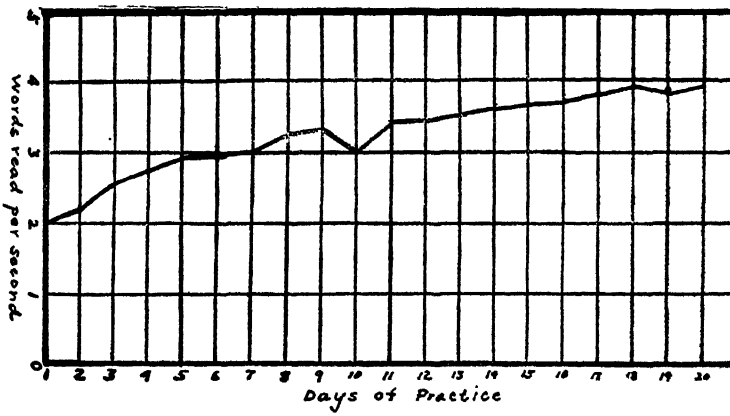


FIG. 20. GRAPH OF DATA PRESENTED IN PROBLEM V

with graphs to show deviation, but they generally greatly complicate the figure.

Whenever it is desirable to present the results of a class on some single performance, such as an examination, there is another method of graphing more satisfactory than those already discussed. In this type of graph the scores are represented on the abscissa or base line and the number of persons reaching a certain score is shown along the ordinate or perpendicular line

PROBLEM VI

Scores of a class of 42 pupils on a test in algebra.

Scores	*No. of Pupils*
50-54	1
55-59	0
60-64	4
65-69	5
70-74	8
75-79	11
80-84	7
85-89	3
90-94	2
95-100	1

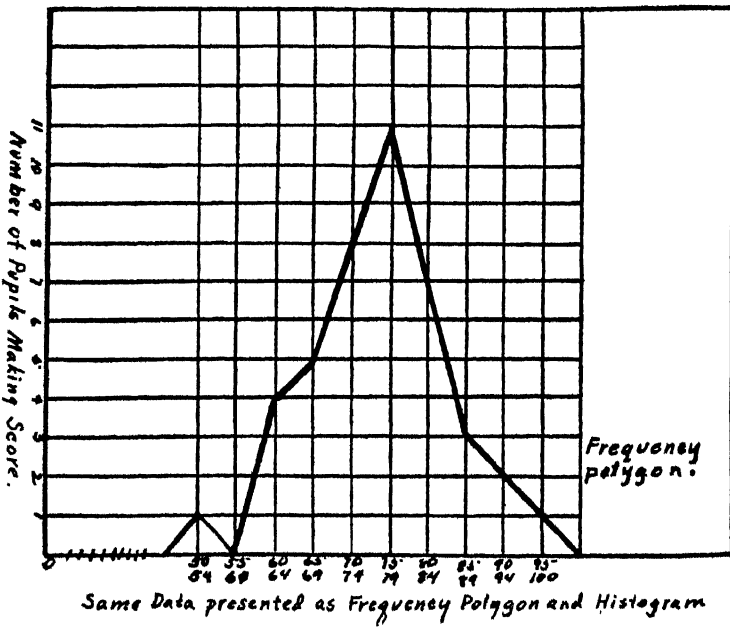


FIG. 21

The graph may be constructed by joining the points on the ordinates indicating the number of cases in each score, by straight lines or by extending horizontal lines from one ordinate to the next at the point representing the number of scores of that value. The former is called

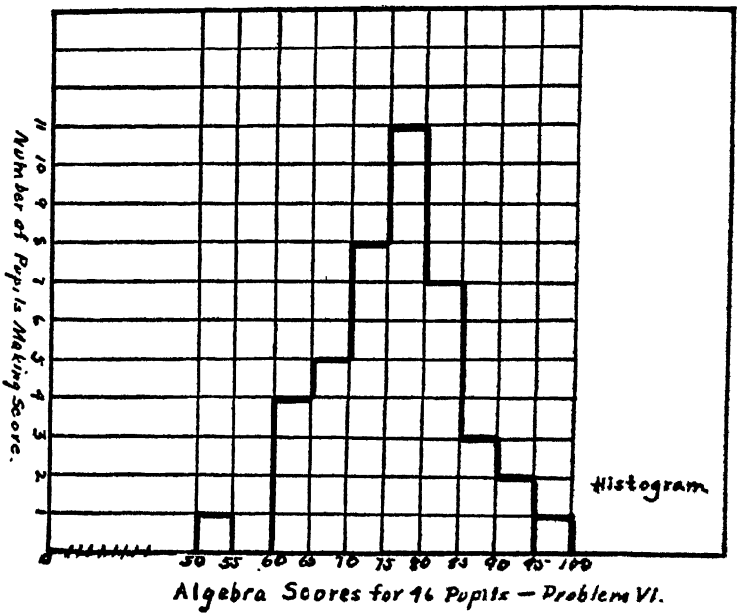


FIG. 22

a frequency polygon and the latter a histogram. There is no essential difference in the two types of graphs. In the histogram there is always the question as to whether the score is represented by the lower edge, the middle or the upper edge of the bar. For example, is the score of 85 to be represented by the beginning, the middle, or the end of the bar? For the sake of uniformity

it is best to represent each score by the bar, beginning with the ordinate representing the score and extending to the next ordinate, that is, make the beginning of the bar represent the score value as illustrated in the graph.

Only the more elementary methods of statistics and graphic presentation have been described in this chapter. The reader desiring a more detailed description of such methods and the more complicated methods and problems are referred to the many good texts in the subject.[3]

[3] The most outstanding texts in statistics as applied to problems of education are: *Statistical Methods Applied to Education* by H. O. Rugg, Houghton Mifflin Co.; *Mental and Social Measurements* by E. L. Thorndike, Teachers College Bureau of Publication, New York; and *Statistical Method* by T. L. Kelley, Macmillan Co.

INDEX